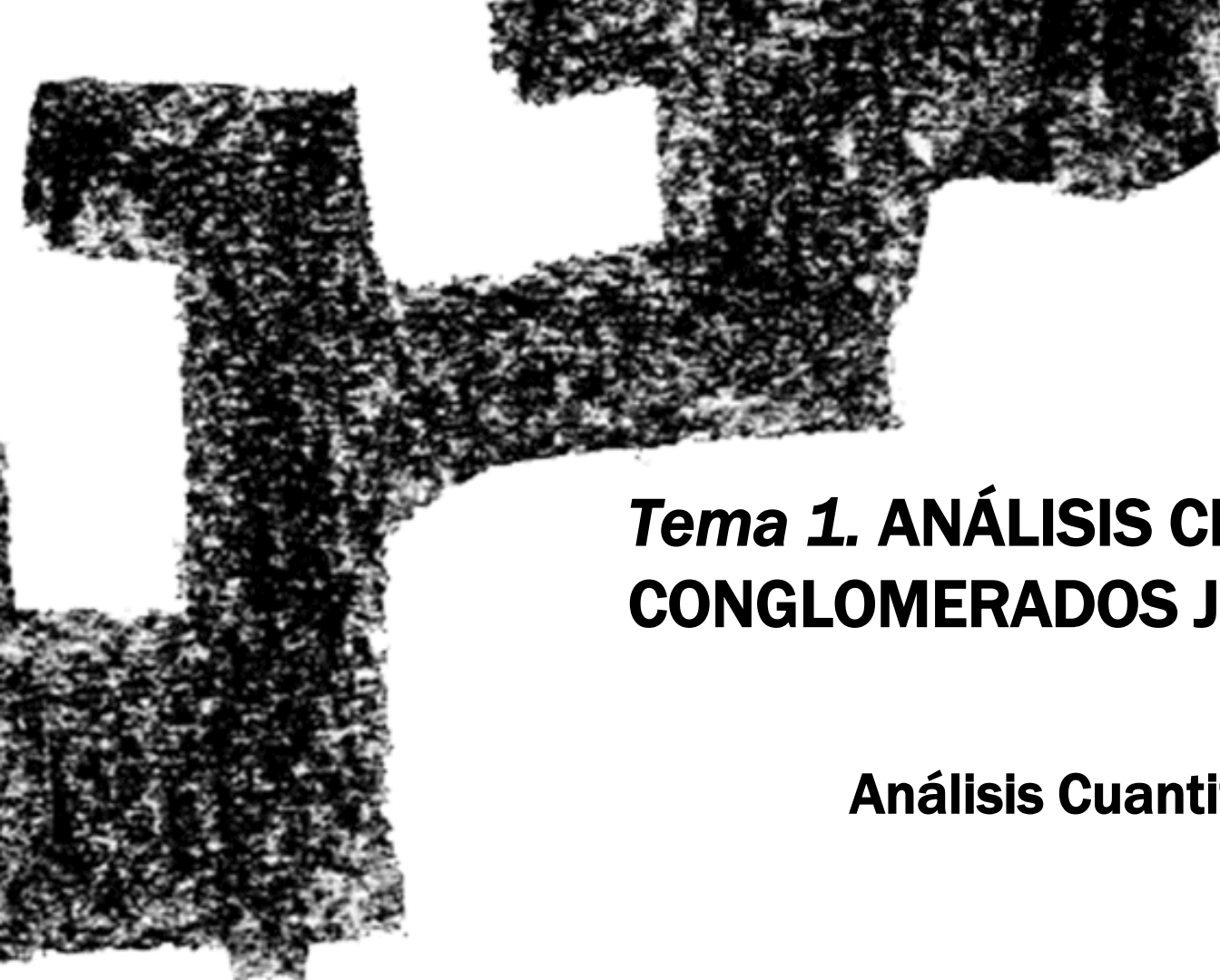




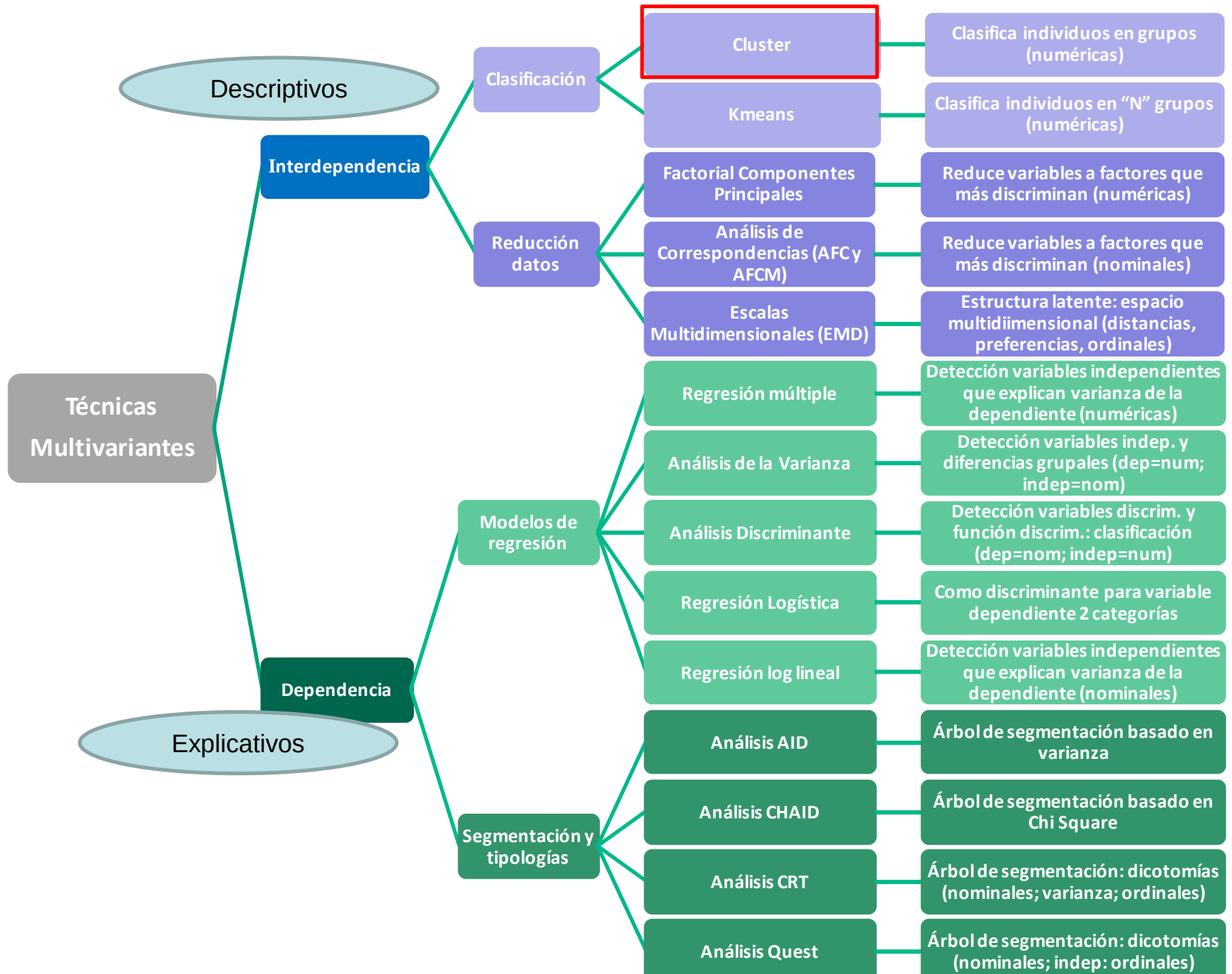
Tema 1. ANÁLISIS CLUSTER **CONGLOMERADOS JERÁRQUICOS**

Análisis Cuantitativo Avanzado



Universidad
del País Vasco

Euskal Herriko
Unibertsitatea





ANÁLISIS CLUSTER CONGLOMERADOS JERÁRQUICOS

- **TIPO:** TECNICA DE INTERDEPENDENCIA, DE CLASIFICACIÓN
- **BUSCA:** PARTICIÓN DE UN COLECTIVO EN SUBGRUPOS (N=?): HOMOGENEIDAD INTRAGRUPPO Y HETEROGENEIDAD INTERGRUPOS
- **NECESITA:** CONJUNTO DE VARIABLES CONTINUAS O DE INTERVALO





FASES ANÁLISIS CLUSTER CONGLOMERADOS JERÁRQUICOS

1. ANÁLISIS DE LA **TABLA DE AGLOMERACIÓN** (O DISOCIACIÓN): DETERMINACIÓN PASOS DE AGRUPACION PERMITIDOS
2. ANÁLISIS **DENDOGRAMA**: VISUALIZACIÓN DE PASOS DE AGRUPACIÓN PERMITIDOS Y DEFINICIÓN NÚMERO SUBGRUPOS DE UN COLECTIVO.

Dendrograma

***** H I E R A R C H I C A L C L U S T E R A N A L Y S I S *****

Dendrogram using Average Linkage (Between Groups)

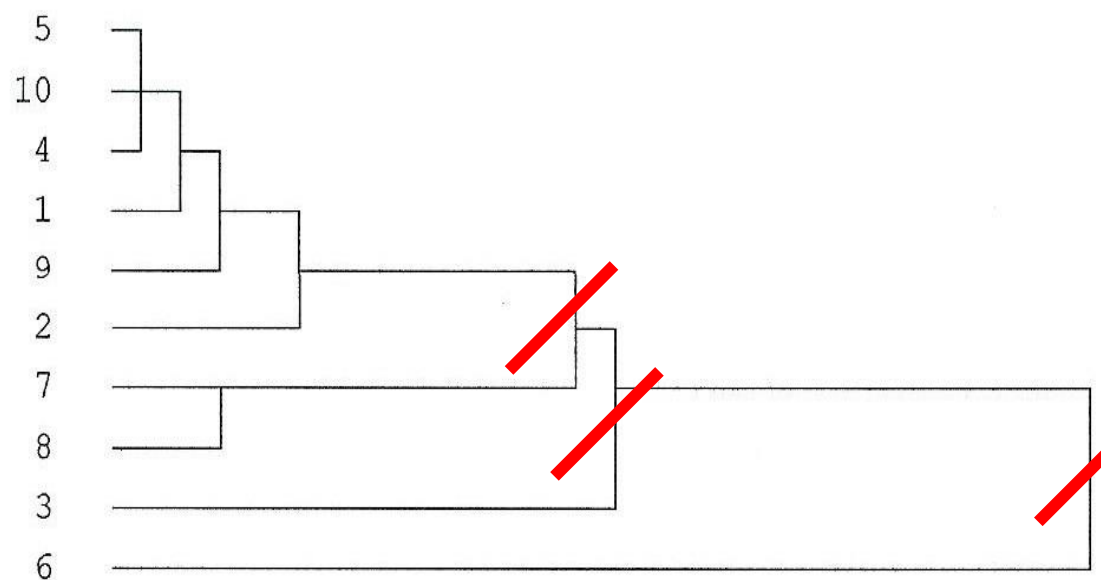
Rescaled Distance Cluster Combine

C A S E

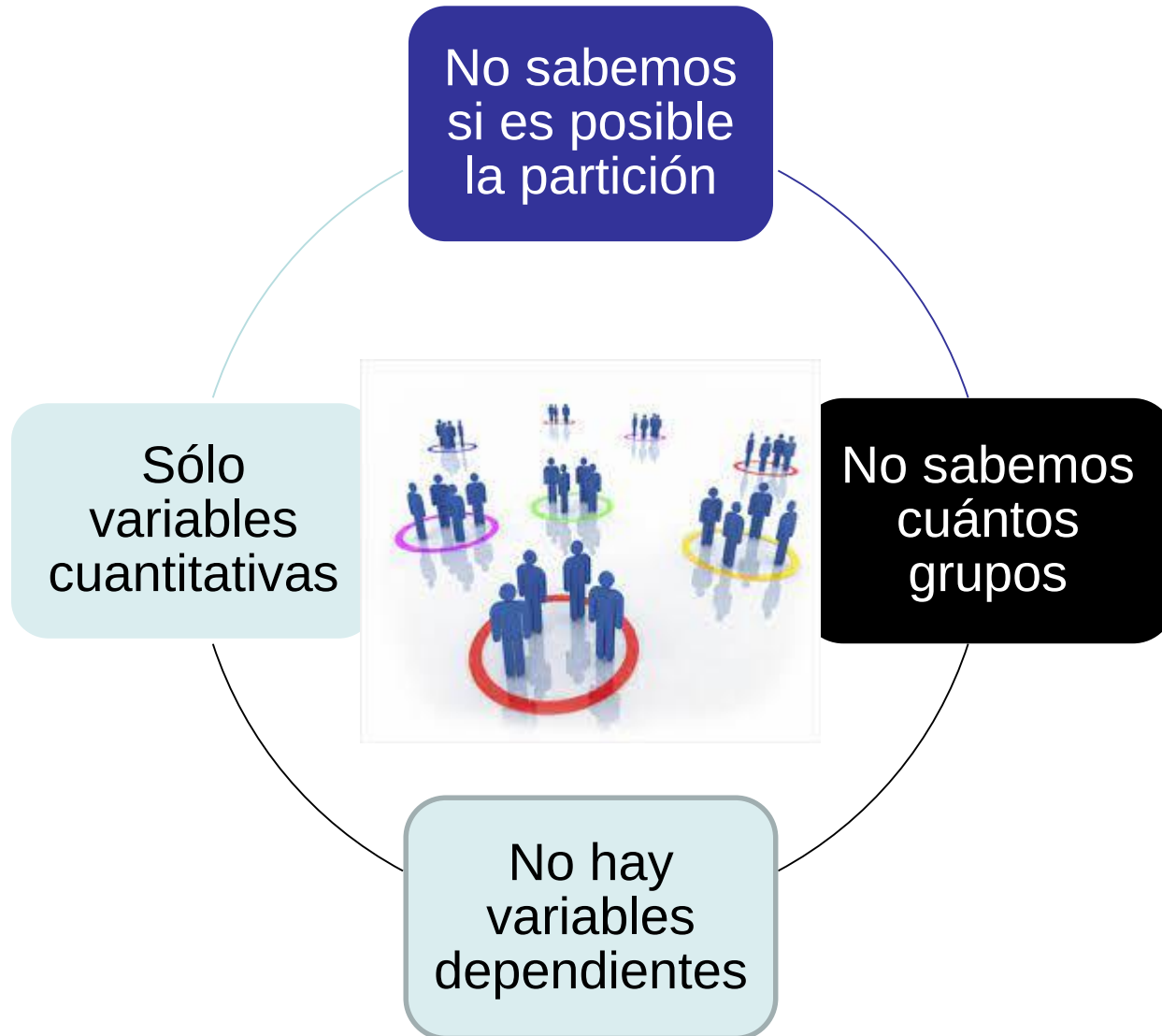
Label

Num 0 5 10 15 20 25
+-----+-----+-----+-----+-----+

PSICOLOGIA
SOCIOLOGIA
PEDAGOGIA
EDUCACION SOCIAL
POLITICAS
TRABAJO SOCIAL
PERIODISMO
PUBLICIDAD
COMUNICACION
PSICOPEDAGOGIA



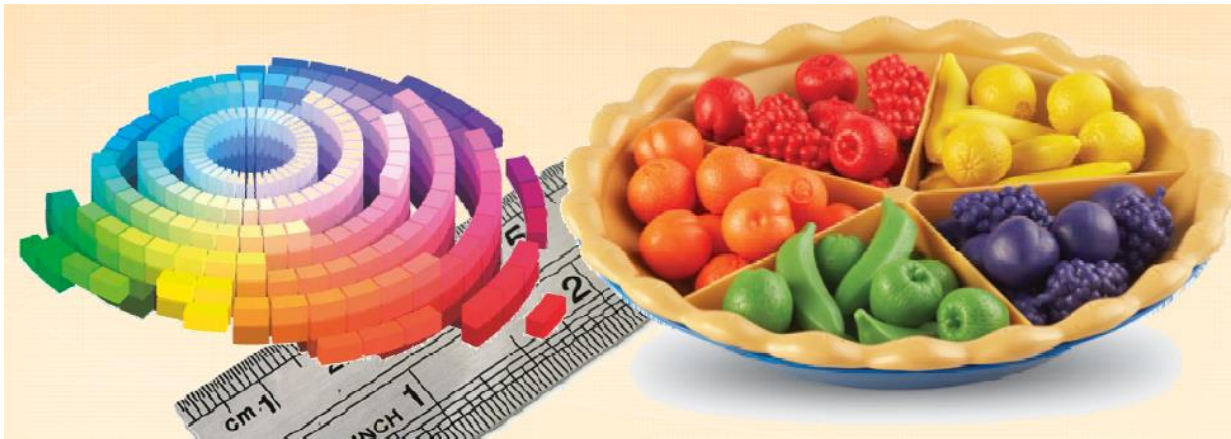
Cluster

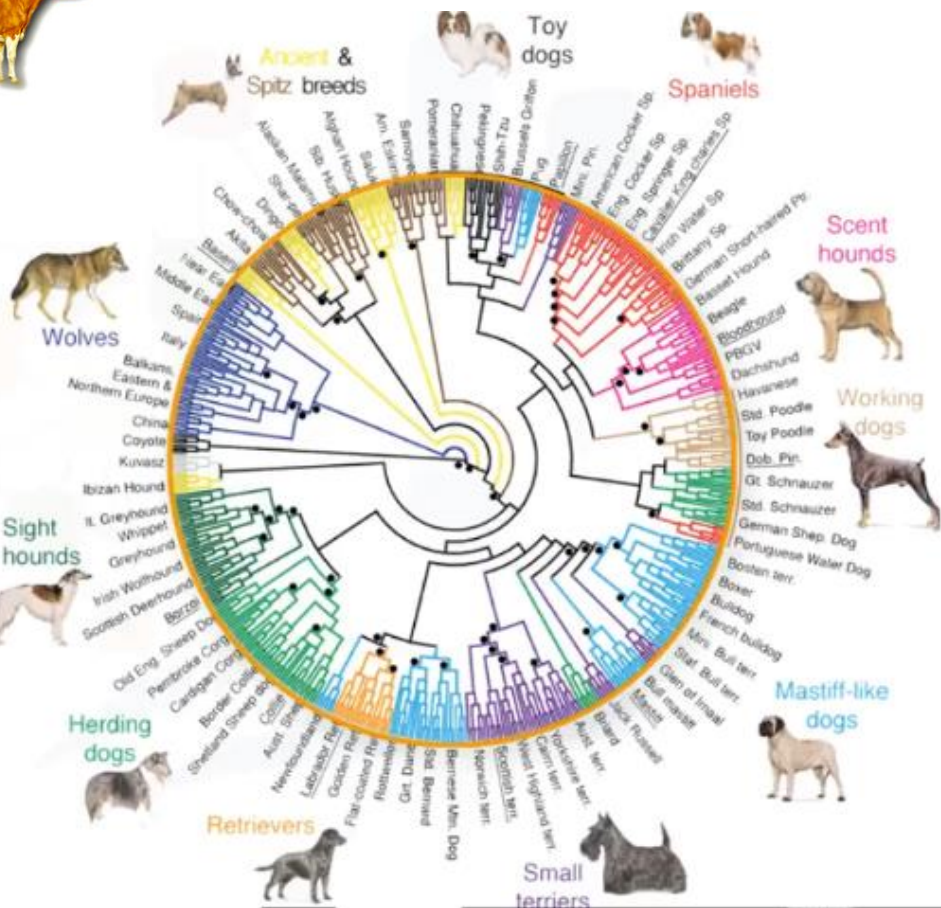


Cluster

Objetivo de la **segmentación**:

- **Variables.** Si el objetivo es resumir el conjunto de variables que se estudian en un número menor de factores, es decir, encontrar la estructura oculta del evento que se analiza. En este caso, el estudio de cluster es similar al de Análisis factorial.
- **Casos.** El objetivo es encontrar clasificaciones o tipologías de casos de grupos homogéneos





Cluster

Es posible clasificar casos en función del color del pelo?

Rubios?

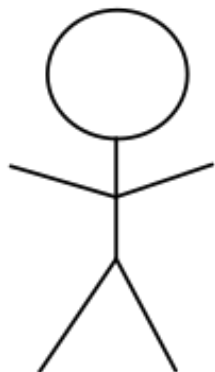
Pelirrojos?

Morenos?

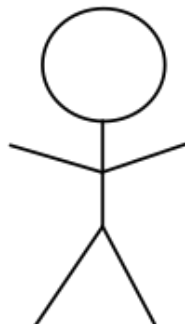




1,90



1,70



1,50





Cluster



Cliente	Edad	Gasto mensual (\$)
A	22	50
B	25	60
C	45	200
D	50	220
E	30	70
F	55	250

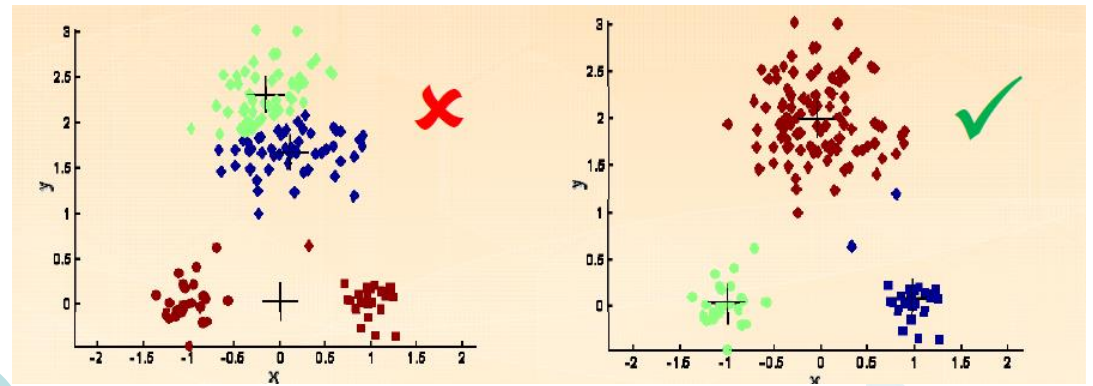
Cluster Objetivos

**Homogeneidad
intragrupo**

- Que los casos que componen cada subgrupo o montón sean lo más parecidos posibles

**Heterogeneidad
intergrupo**

- Hacer los grupos sean lo más diferentes posible





Cluster

Preguntas a considerar...

- ¿**Cómo medir la proximidad**? El objetivo es agrupar los casos que son similares en varios grupos. Se pueden utilizar diferentes métodos (en el espacio bidimensional la **distancia** es la más común, pero también se pueden utilizar correlaciones y medidas de asociación)
- ¿**Cómo crear grupos**? Cualquiera que sea la medida de proximidad, el procedimiento debe unir elementos que sean similares
- ¿**Cuántos grupos formar**? Existen varios métodos para medir y valorar la homogeneidad intragrupo, pero el objetivo final es buscar el equilibrio entre las distancias dentro del conjunto y entre los grupos.

Cluster

Criterios

- Sea cual sea el proceso de aglomeración, hay varios criterios para formar grupos
- Todas ellas se basan en una matriz de distancias o de proximidad

Matriz de proximidades

Caso	Distancia euclídea al cuadrado									
	1:Gran Bretaña	2:Finlandia	3:Belgica	4:Alemania	5:Israel	6:Italia	7:Holanda	8:España	9:Suecia	10:Suiza
1:Gran Bretaña	,000	2386,840	2787,430	990,570	4191,120	3382,210	3177,960	3193,940	3476,290	3641,000
2:Finlandia	2386,840	,000	1919,730	1794,410	1836,860	3792,050	1357,700	1521,340	2865,370	2466,060
3:Belgica	2787,430	1919,730	,000	2458,740	3399,490	2930,000	744,450	1405,930	3151,900	485,290
4:Alemania	990,570	1794,410	2458,740	,000	3722,070	3792,680	2215,670	3367,150	2462,760	3042,330
5:Israel	4191,120	1836,860	3399,490	3722,070	,000	4799,850	3799,140	2697,940	3448,890	4271,180
6:Italia	3382,210	3792,050	2930,000	3792,680	4799,850	,000	4483,630	4578,910	1687,500	4600,550
7:Holanda	3177,960	1357,700	744,450	2215,670	3799,140	4483,630	,000	1241,220	4074,650	1049,300
8:España	3193,940	1521,340	1405,930	3367,150	2697,940	4578,910	1241,220	,000	4778,630	1859,420
9:Suecia	3476,290	2865,370	3151,900	2462,760	3448,890	1687,500	4074,650	4778,630	,000	4030,170
10:Suiza	3641,000	2466,060	485,290	3042,330	4271,180	4600,550	1049,300	1859,420	4030,170	,000

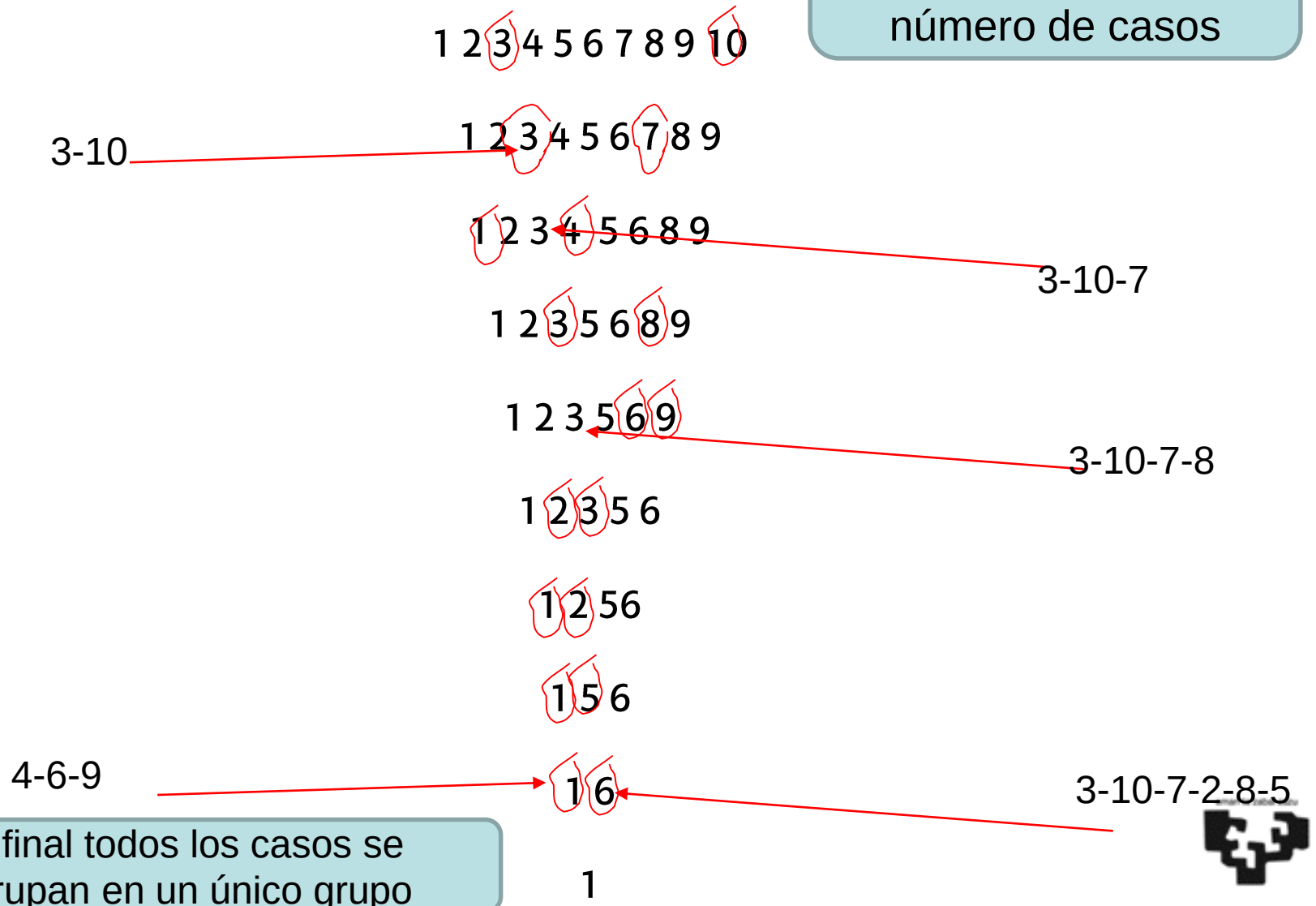
Esto es una matriz de disimilaridad.





Cluster

Al inicio el número de grupos es igual al número de casos



Cluster

Criterios

- A la hora de elegir el criterio, se utilizan medidas normales de **distancia** o proximidad: distancia euclídea, cuadrado de distancia euclídea, vector coseno, etc.

Medida	
☒ Intervalo:	Distancia euclídea al cuadrado
	Distancia euclídea
	Distancia euclídea al cuadrado
☐ Recuentos:	Coseno
	Correlación de Pearson
☐ Binaria:	Chebychev
	Bloque
	Minkowski
Transformar valor	Personalizada

- En cualquier caso, no hay una única solución y no hay un método o medida infalible
- Los resultados pueden variar según el método utilizado
- Como siempre, el rigor, el conocimiento del problema y la experiencia del investigador/a guiarán la elección del método y resultado adecuado



Cluster Hurbiltasunen neurria

Distancia euclídea

- ***Distancia euclídea***. Medida de disimilaridad utilizada por defecto para datos de intervalo. Raíz cuadrada de la suma de los cuadrados de las diferencias entre los valores de las variables:

$$EUCLID(X, Y) = \sqrt{\sum_i (X_i - Y_i)^2}$$

Distancia euclídea al cuadrado

- ***Distancia euclídea al cuadrado***. Medida de disimilaridad. Suma de los cuadrados de las diferencias entre los valores de las variables:

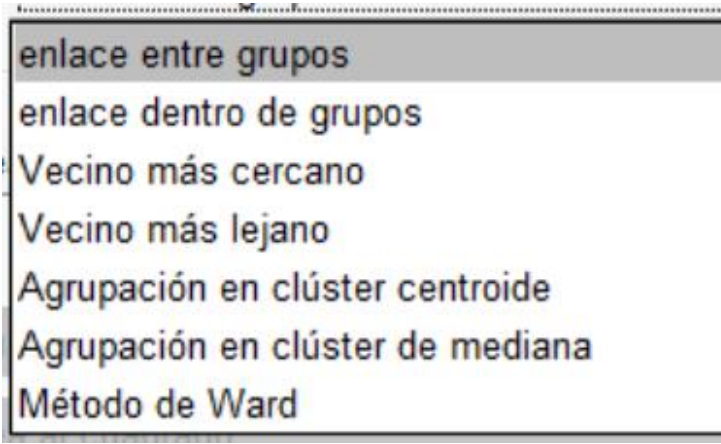
$$SEUCLID(X, Y) = \sum_i (X_i - Y_i)^2$$



Cluster

Métodos de aglomeración

- Enlace entre grupos (UPGMA o **BAVERAGE**): Inter-grupos
- Enlace dentro de grupos (**WAVERAGE**): Intra-grupos
- Vecino más cercano (**Single**)
- Vecino más lejano (**Complete**)
- Agrupación en clúster centroide (**Centroid**)
- Agrupación en clúster de mediana (**Median**)
- Método **Ward**



enlace entre grupos
enlace dentro de grupos
Vecino más cercano
Vecino más lejano
Agrupación en clúster centroide
Agrupación en clúster de mediana
Método de Ward

Cluster

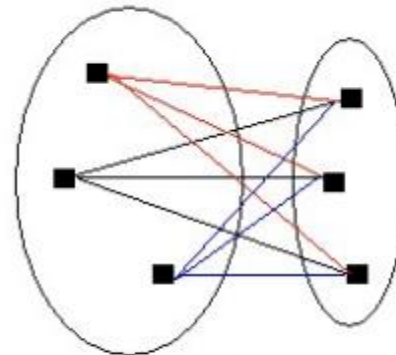
Enlace entre grupos (UPGMA o BAVERAGE)

El método del promedio entregrupos (UPGMA o baverage) define la distancia entre dos conglomerados como el promedio de las distancias entre todos los pares de individuos que se puedan formar entre los dos grupos. Suele ser uno de los métodos preferidos por tener en cuenta a todos los miembros de los grupos, no sólo los más cercanos o lejanos.

- Promedio de grupo

$$d(A, B) = \frac{1}{n_A \cdot n_B} \sum_{i \in A, j \in B} d(i, j)$$

conglomerados más robustos



promedio de grupo

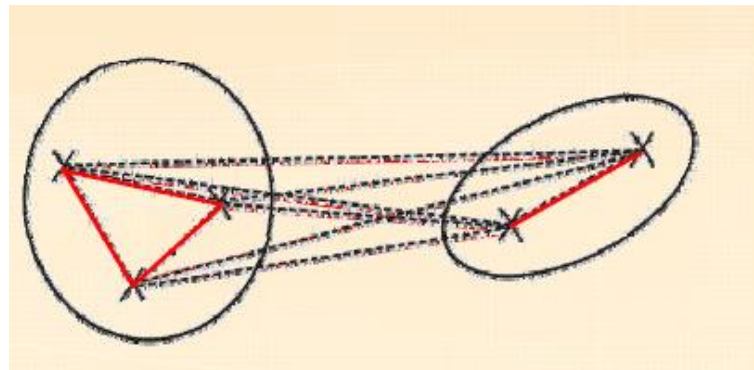
$$d(A, B) = \frac{1}{n_A \cdot n_B} \sum_{i \in A, j \in B} d(i, j)$$



Cluster

Enlace dentro de grupos WAVERAGE

- El método de la media ponderada es una variación del anterior (waverage) y combina los grupos de tal forma que la distancia promedio entre todos los casos del cluster resultante sea lo más pequeña posible



Cluster

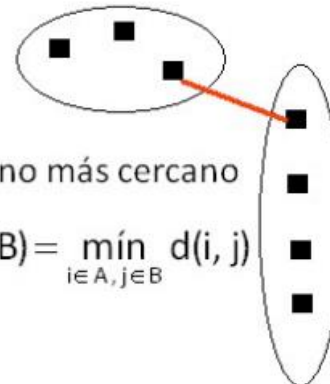
Vecino más cercano (SINGLE)

- El método de las distancias mínimas (single) es uno de los métodos más sencillos. Consiste en agrupar los casos que tienen menor distancia o mayor similitud: la distancia entre dos grupos es la distancia de sus puntos más cercanos.
- Detecta bien los casos atípicos (outliers)

- Vecino más cercano:

$$d(A, B) = \min_{i \in A, j \in B} d(i, j)$$

conglomerados más alargados



$$d(A, B) = \min_{i \in A, j \in B} d(i, j)$$

Cluster

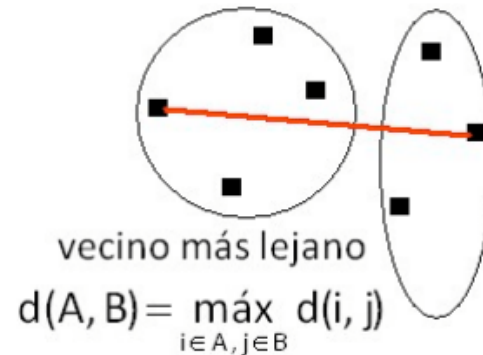
Vecino más lejano (COMPLETE)

- Toma como criterio de agrupación la distancia entre los puntos o grupos más alejados

- Vecino más lejano

$$d(A, B) = \max_{i \in A, j \in B} d(i, j)$$

conglomerados más esféricos



- No funciona bien si hay outliers

Cluster

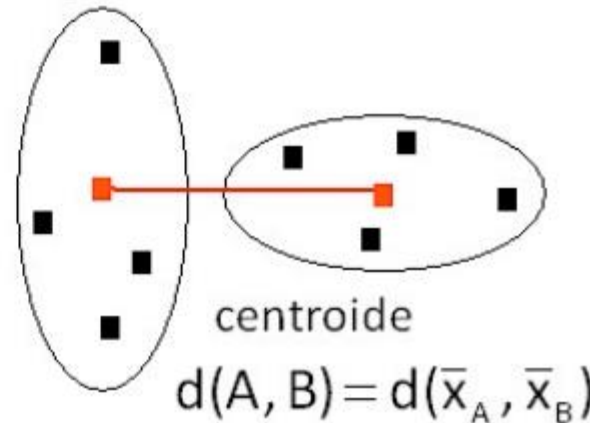
METODO CENTROIDE (CENTROID)

- El método centroeide (centroid) calcula la proximidad entre dos grupos como la distancia entre sus medias para todas las variables.
- Se debe de utilizar como unidad de medida la distancia **euclídea al cuadrado**.

- Centroeide (centro gravedad)

$$d(A, B) = d(\bar{x}_A, \bar{x}_B)$$

conglomerados más robustos



- **Problemas:**
 - No funciona bien con los outliers



Cluster

MÉTODO DE LA MEDIANA (MEDIAN)

- El método de la mediana (median) se basa en agrupar cuando los valores de la mediana de cada uno de los grupos son próximos.
- Se debe de utilizar la distancia **euclídea al cuadrado**.
- No tiene en cuenta el número de casos

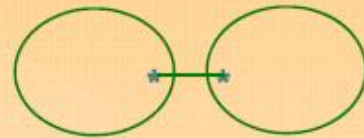


Cluster

MÉTODO WARD

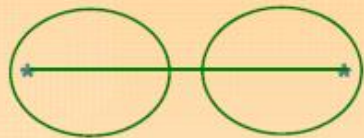
- El método de Ward (Ward) calcula las medias de cada grupo.
- Calcula la distancia euclídea al cuadrado de cada caso con la media de su grupo, sumando en cada grupo las distancias de todos sus componentes.
- En cada paso los dos grupos que se unen son aquéllos que proporcionan el menor incremento en la suma total del cuadrado de la distancia intragrupo.
- Se debe de usar el cuadrado de la **distancia euclídea**.
- Crea grupos más homogéneos y equilibrados de tamaño

Cluster Métodos



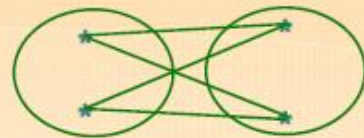
Vinculación única:
Mínima distancia

Single



Vinculación completa:
Máxima distancia

Complete



Vinculación media
(inter-grupos):
Distancia media

Baverage



Método de centroide:
Distancia entre centro

Centroide



Método de Ward:
*Minimización de la
varianza intragrupos*

Ward



Cluster

¿Qué par de elementos se unirán primero?

A*

*B

*G

C *

*D

H*

*E

Cluster

Depende de la medida de distancia que utilizemos, pero supongamos que A-B y D-E. Surgen así dos conglomerados.

A* — B*

*G

C *

*D

*E

H*

Cluster

El siguiente punto más próximo a estos dos conglomerados es C. ¿A dónde será adscrito?

A* ————— *B

*G

C *

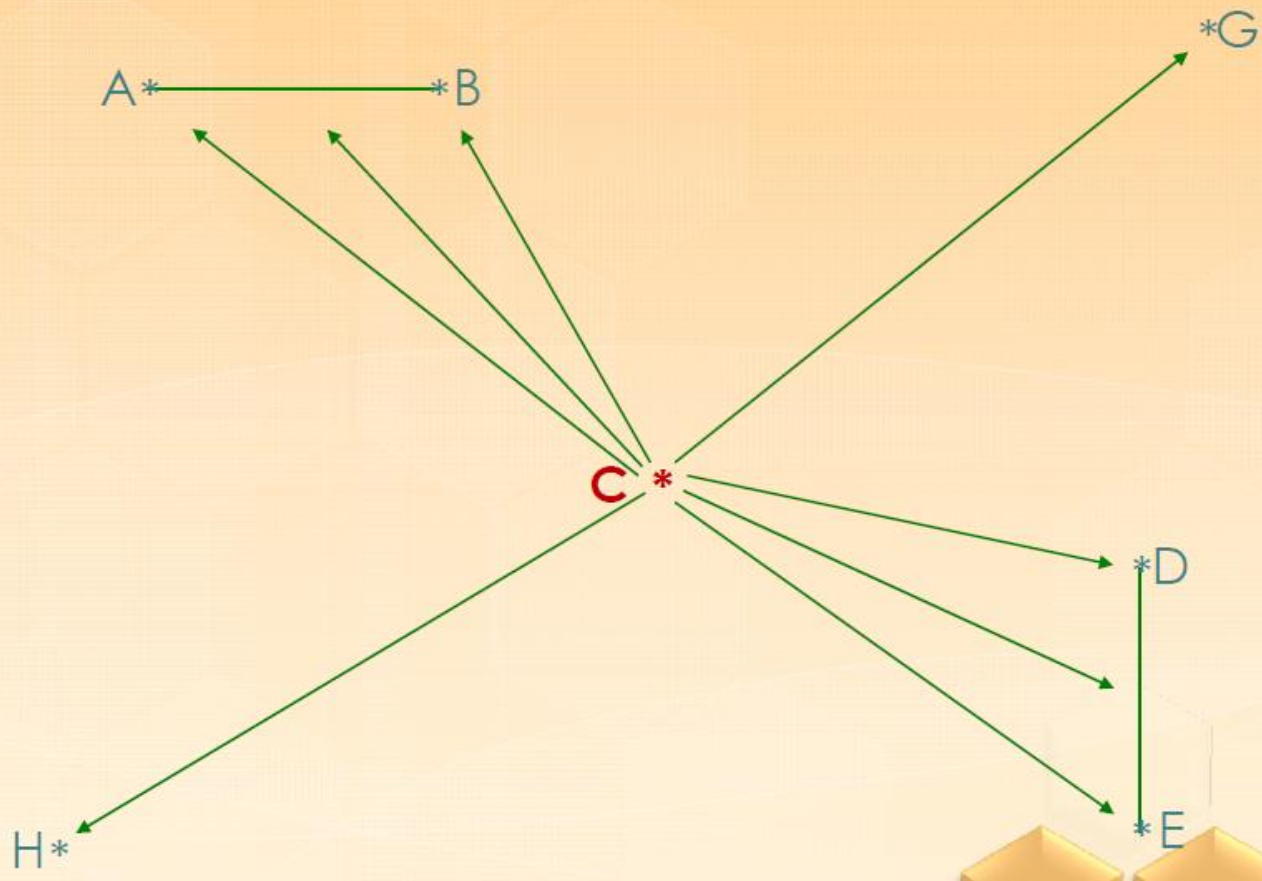
H*

*D

*E

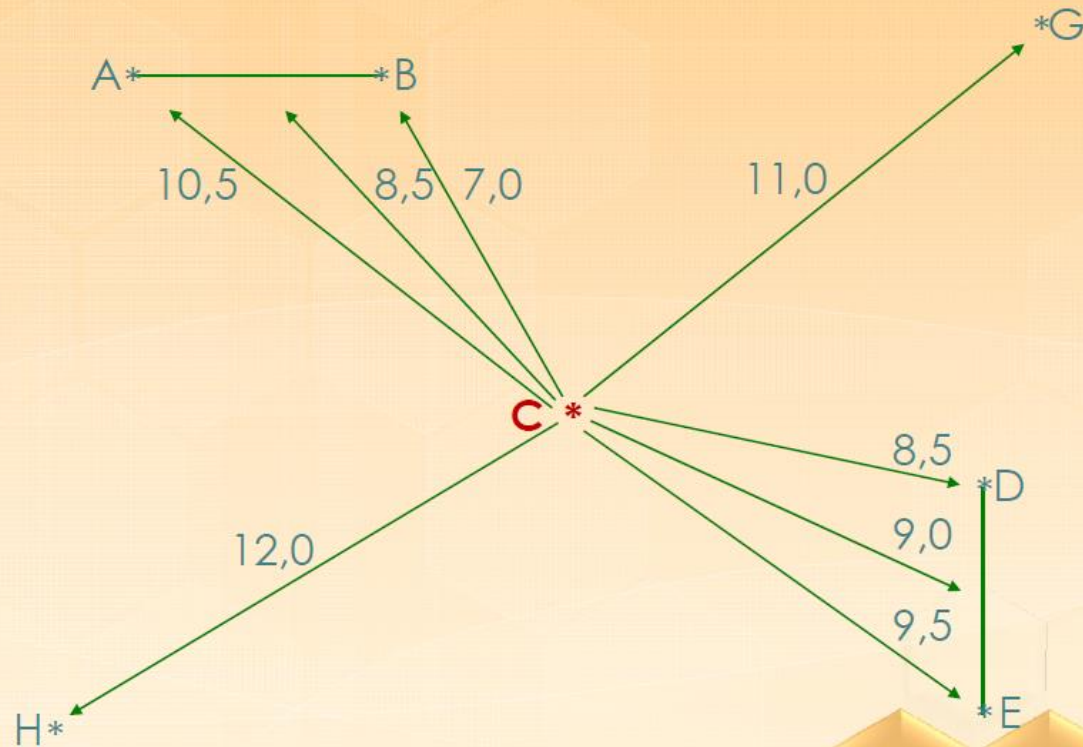
Cluster

A dónde se encuentre más próximo (distancia euclídea), pero... EXISTEN VARIAS POSIBILIDADES (métodos de agrupamiento).



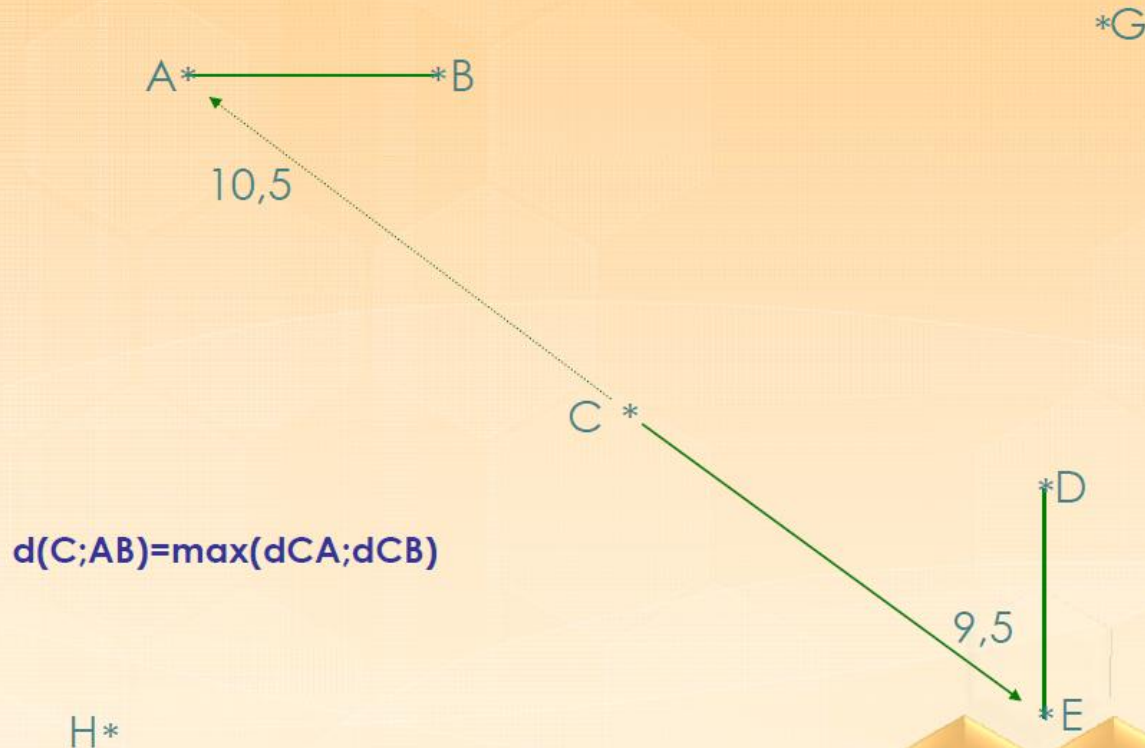
Cluster

A dónde se encuentre más próximo (distancia euclídea), pero... EXISTEN VARIAS POSIBILIDADES (métodos de agrupamiento):



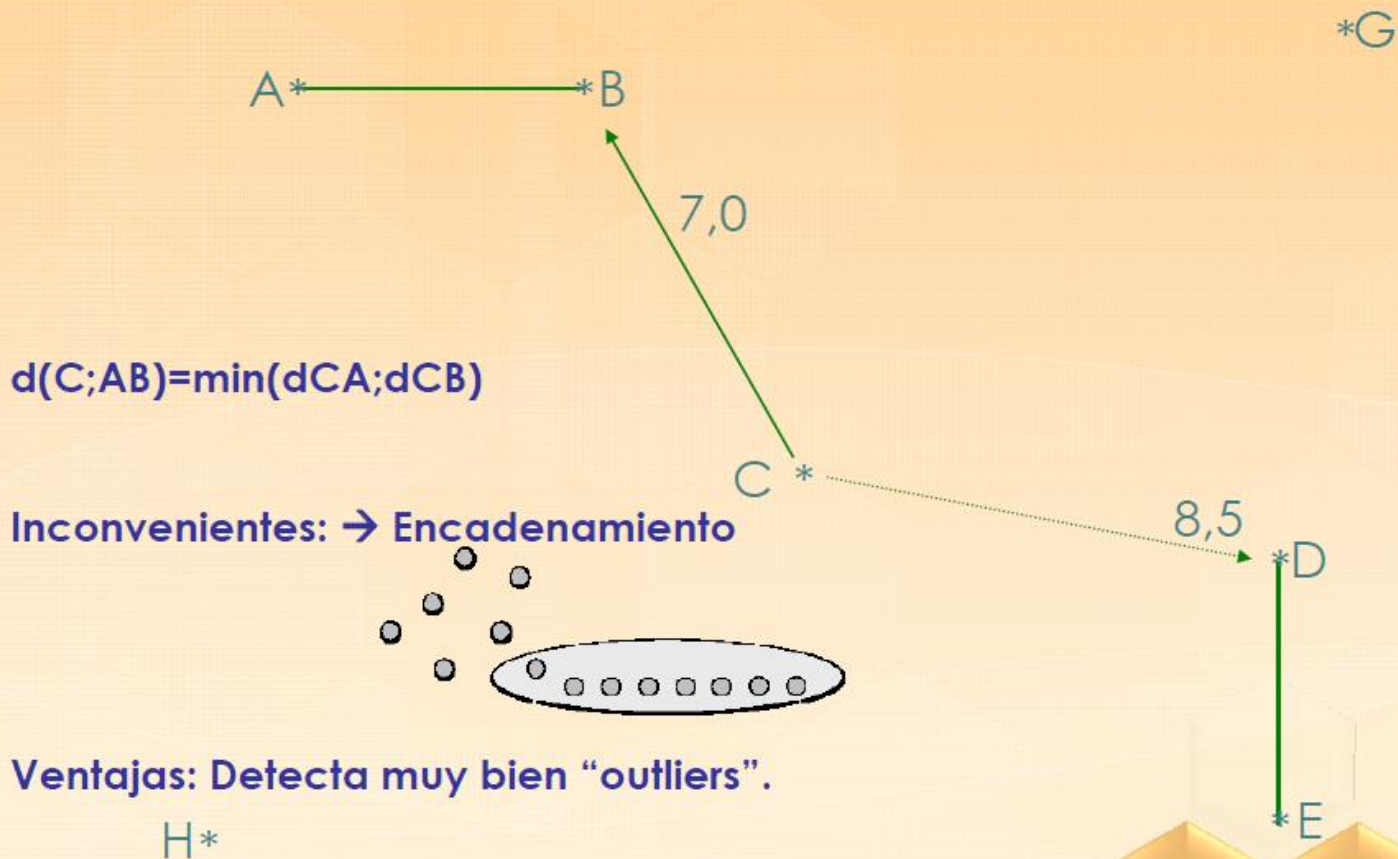
Cluster

VECINO MÁS LEJANO (COMPLETE LINKAGE): Adscribe C al conglomerado cuya distancia el caso más alejado de ese conglomerado es menor.



Cluster

VECINO MÁS PRÓXIMO (SINGLE LINKAGE): Adscribe C al conglomerado cuyo distancia al caso más cercano de ese conglomerado es menor.



Cluster

MÉTODOS BASADOS EN PUNTUACIONES MEDIAS: Adscribe C al conglomerado cuya promedio de distancias es menor.

- Vinculación Intragrupos
- Vinculación intergrupos
- Mediana
- Centroide
- Ward

Tienden a obtener conglomerados de tamaño similar.

Preferibles al resto de jerárquicos, especialmente WARD y VINCULACIÓN INTERGRUPOS.





Cluster Pasos

- 1) Elegir las variables:** Aquellas que realmente explican el fenómeno
- 2) Elegir la medida de proximidad:** Las más usadas distancia euclídea y euclídea al cuadrado
- 3) Elegir el método de agrupación**
- 4) Medida de las variables:** La escala tiene que ser igual



Cluster jerárquico

INTERPRETACIÓN

- **Información numérica**
 - 1) Historial de Conglomeración
- **Información gráfica**
 - 2) Dendograma
 - 3) Diagrama de Témpanos

La iterpretación tiene 3 fases

Cluster jerárquico

0. Variables y métodos

Porcentaje de jóvenes que tienen diferentes aparatos en sus habitaciones, por país

Resumen de procesamiento de casos^{a,b}

Válido		Casos Perdidos		Total	
N	Porcentaje	N	Porcentaje	N	Porcentaje
10	100,0	0	,0	10	100,0

a. Distancia euclídea al cuadrado utilizada

b. Enlace promedio (entre grupos)

Método

Distancia elegida

Número de casos
Número de grupos N-1



Cluster jerárquico

1. Historial de Conglomeración:

Los casos que se van agrupando en cada paso (países)

Historial de conglomeración

Etapa	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapa siguiente
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	3	10	485,290	0	0	2
2	3	7	896,875	1	0	4
3	1	4	990,570	0	0	7
4	3	8	1502,190	2	0	6
5	6	9	1687,500	0	0	9
6	2	3	1816,208	0	4	7
7	1	2	2806,547	3	6	8
8	1	5	3416,829	7	0	9
9	1	6	3790,534	8	5	0

La medida de proximidad

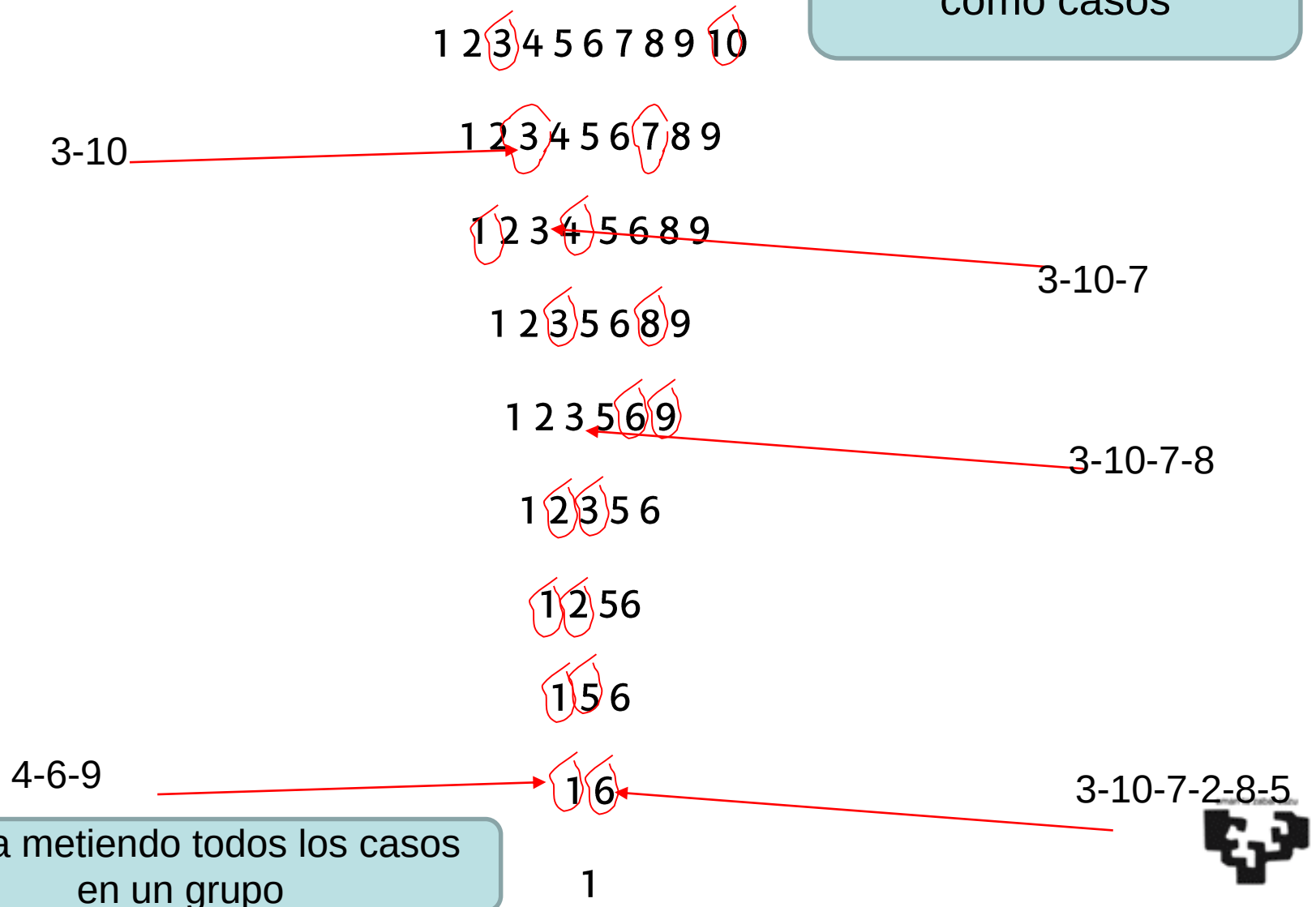
Para marcar el salto





Cluster

Parte de tantos grupos
como casos



Cluster jerárquico

1. Historial de Conglomeración

En la etapa 6 se ha acumulado
el 45,5% del total de
diferencias

Historial de conglomeración						
Etapa	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapa siguiente
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	3	10	485,290	411	0	2
2	3	7	896,875	96	0	4
3	1	4	990,570	512	0	7
4	3	8	1502,190	185	0	6
5	6	9	1687,500	129	4	7
6	2	3	1816,208	990	6	8
7	1	2	2806,547		0	9
8	1	5	3416,829		5	0
9	1	6	3790,534			

Parar

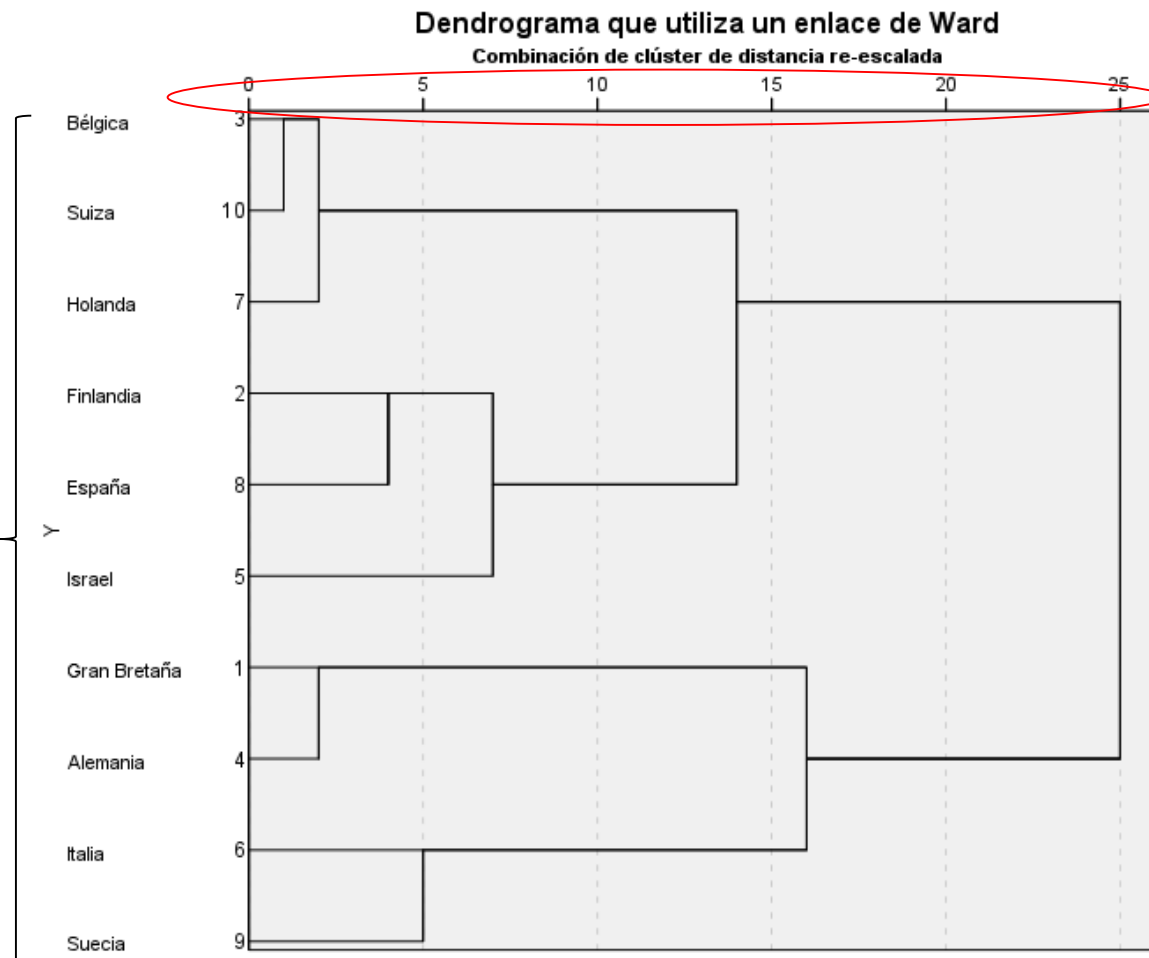
Número de grupos=Número de casos- Número de pasos permitidos
Ejemplo: 10-6=4



Cluster jerárquico

2. Dendrograma:

Casos



Estandarizado

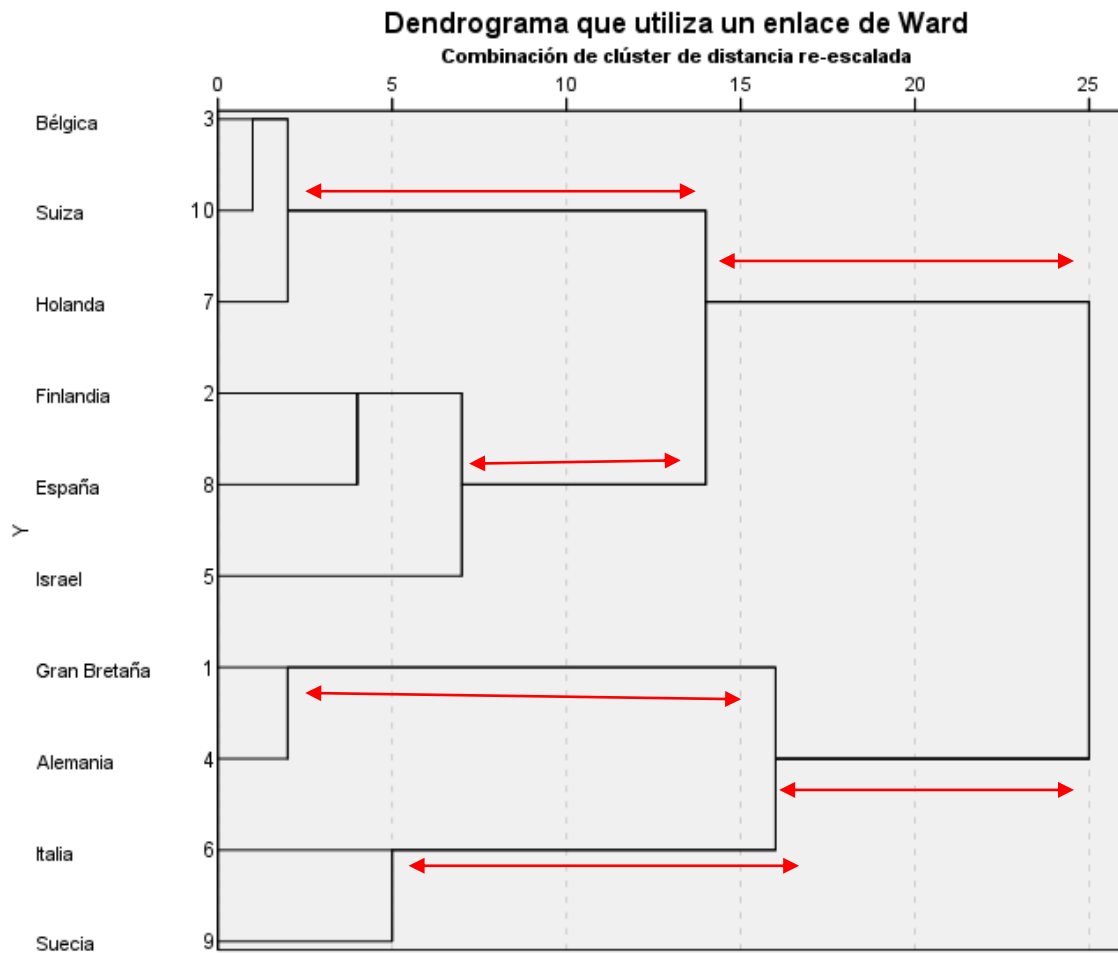
A la izquierda:
elementos parecidos

A la derecha: los
elementos más
distantes



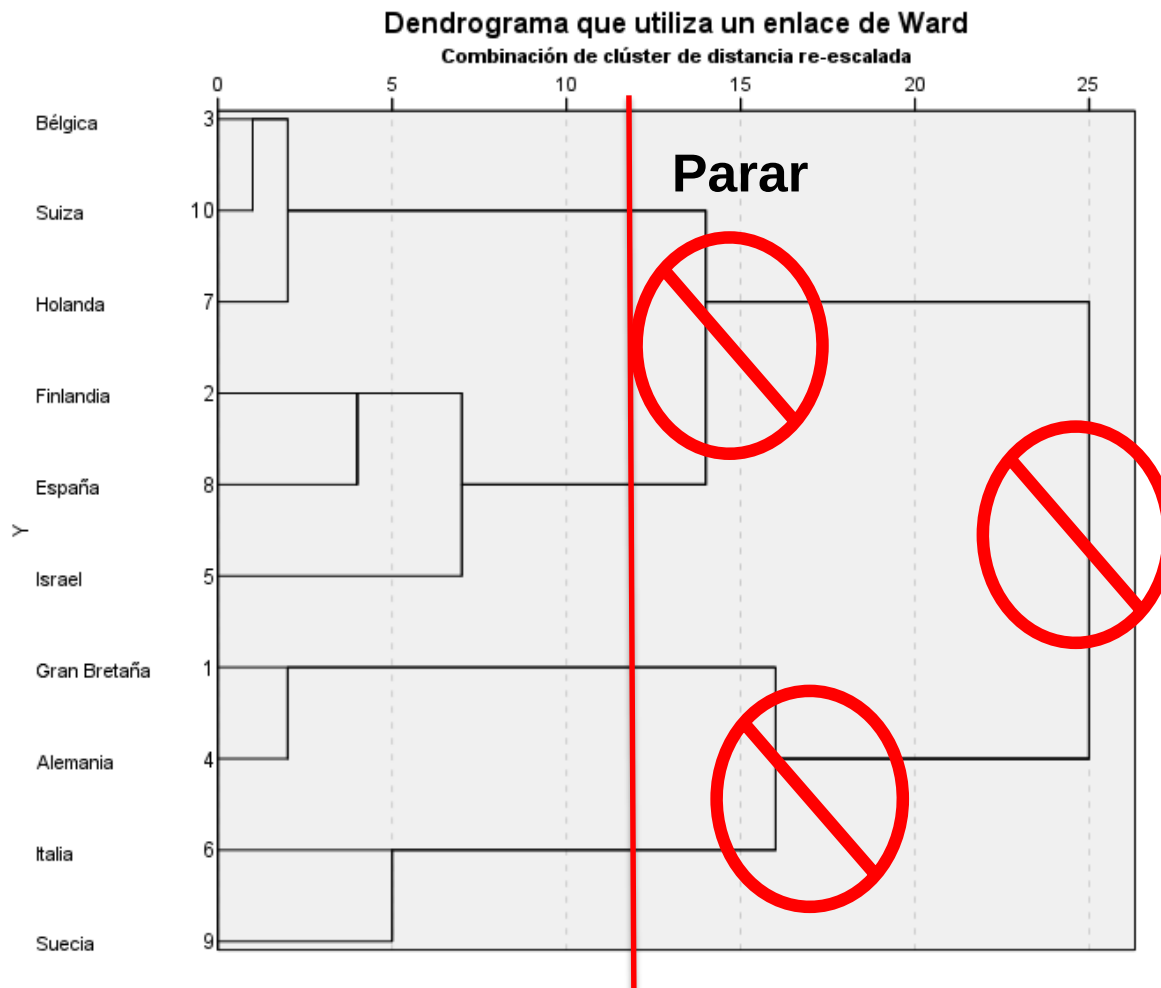
Cluster jerárquico

2. Dendrograma:



Cluster jerárquico

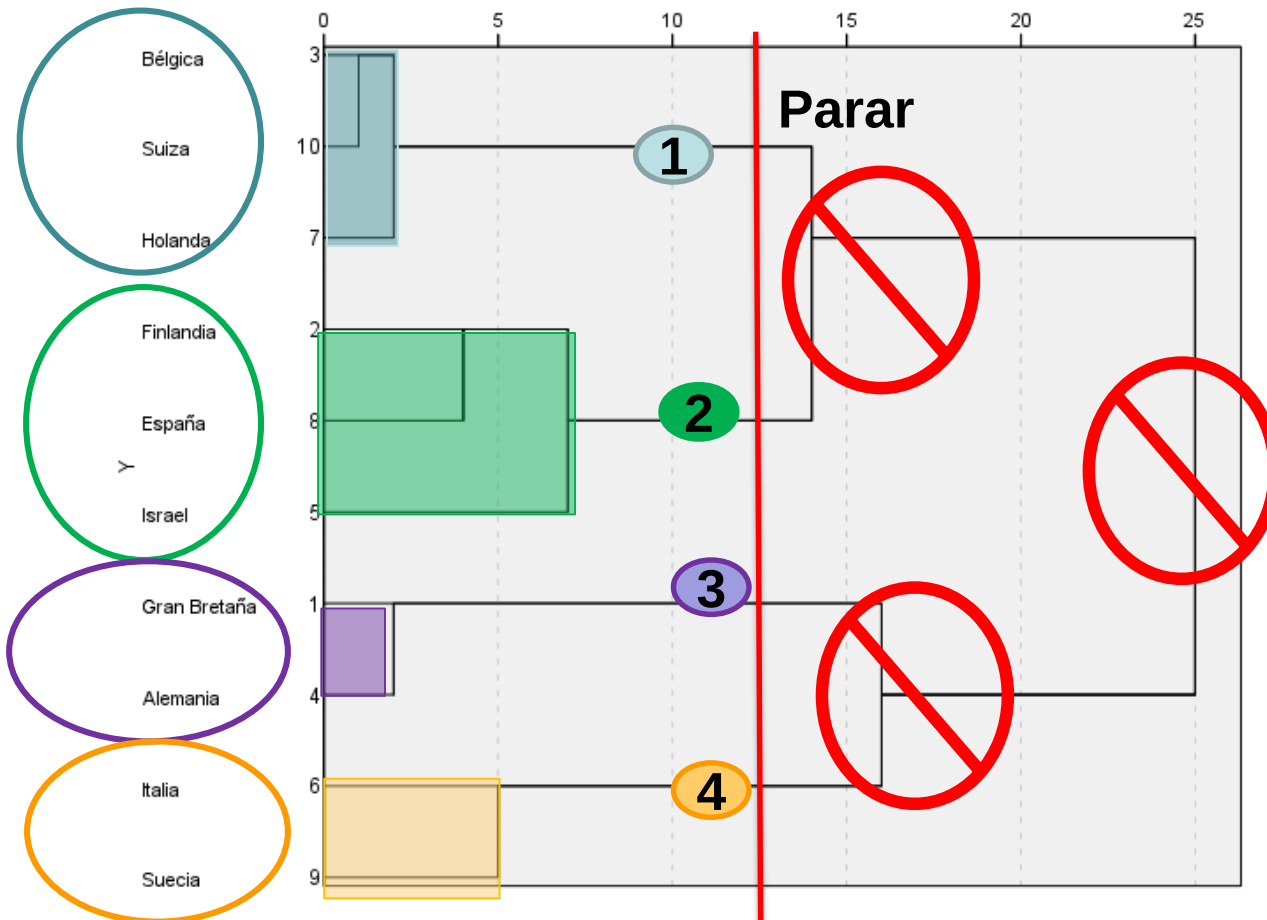
2. Dendrograma:



Cluster jerárquico

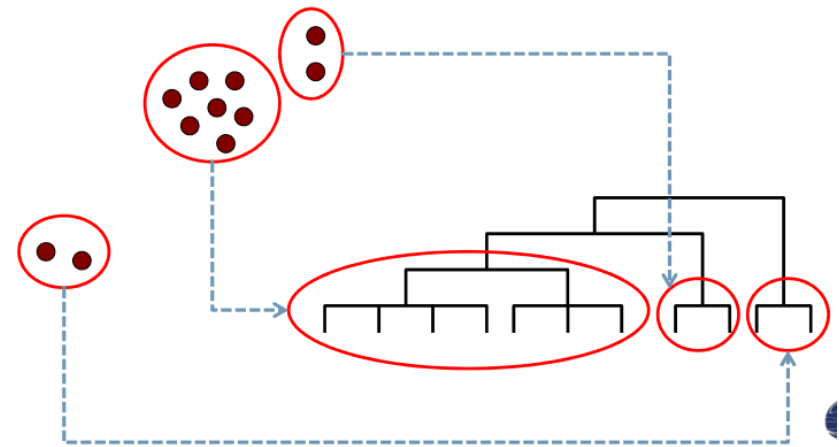
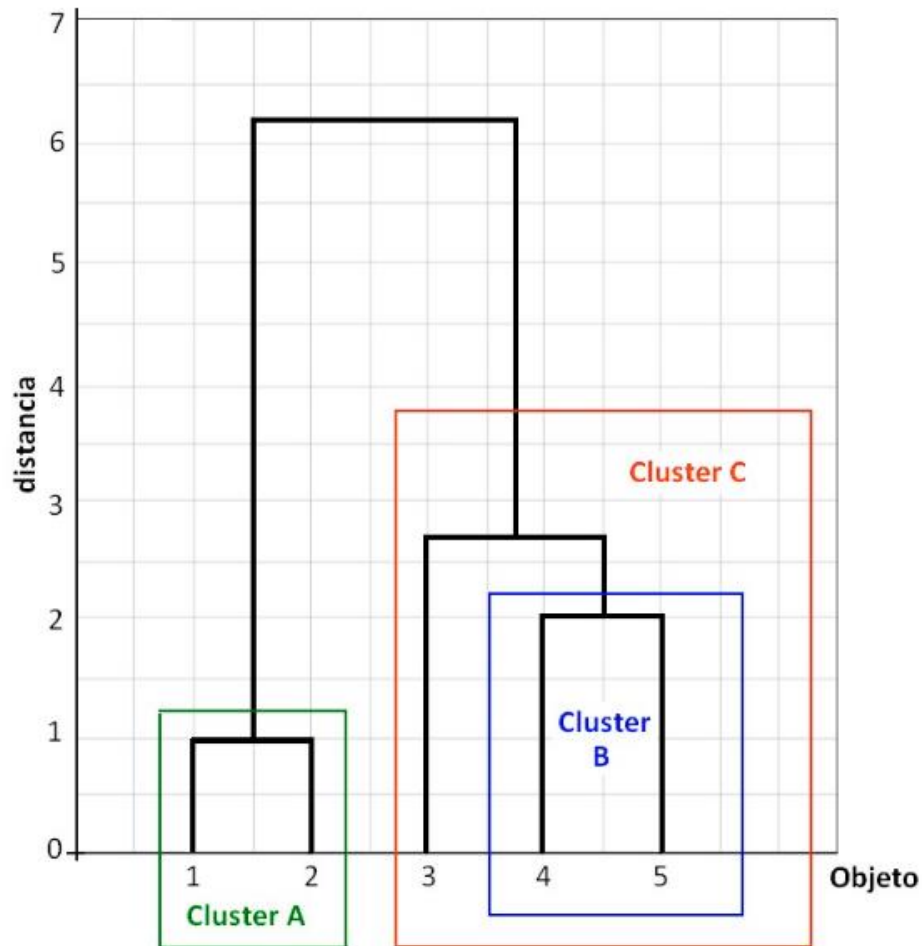
2. Dendrograma:

Dendrograma que utiliza un enlace de Ward
Combinación de clúster de distancia re-escalada



Cluster jerárquico

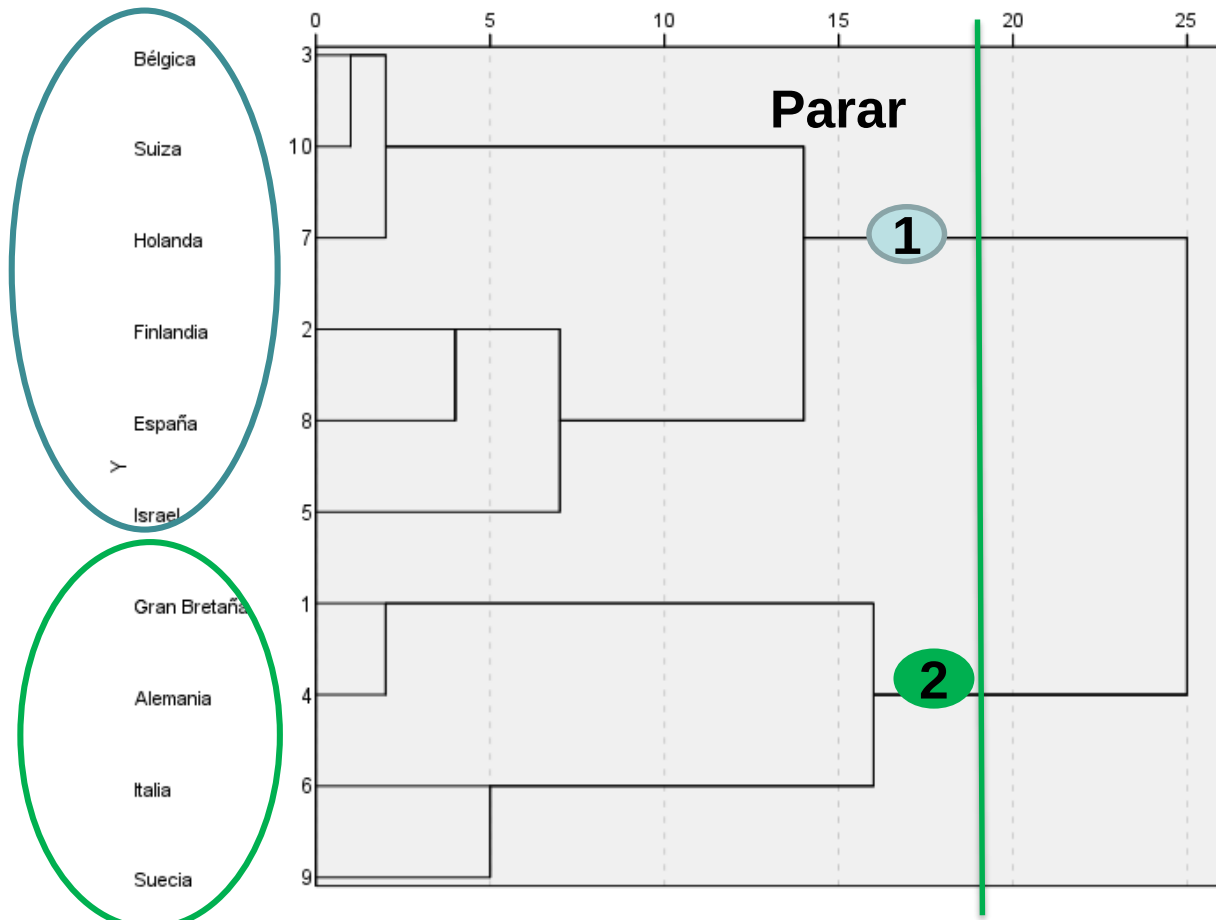
2. Dendograma:



Cluster jerárquico

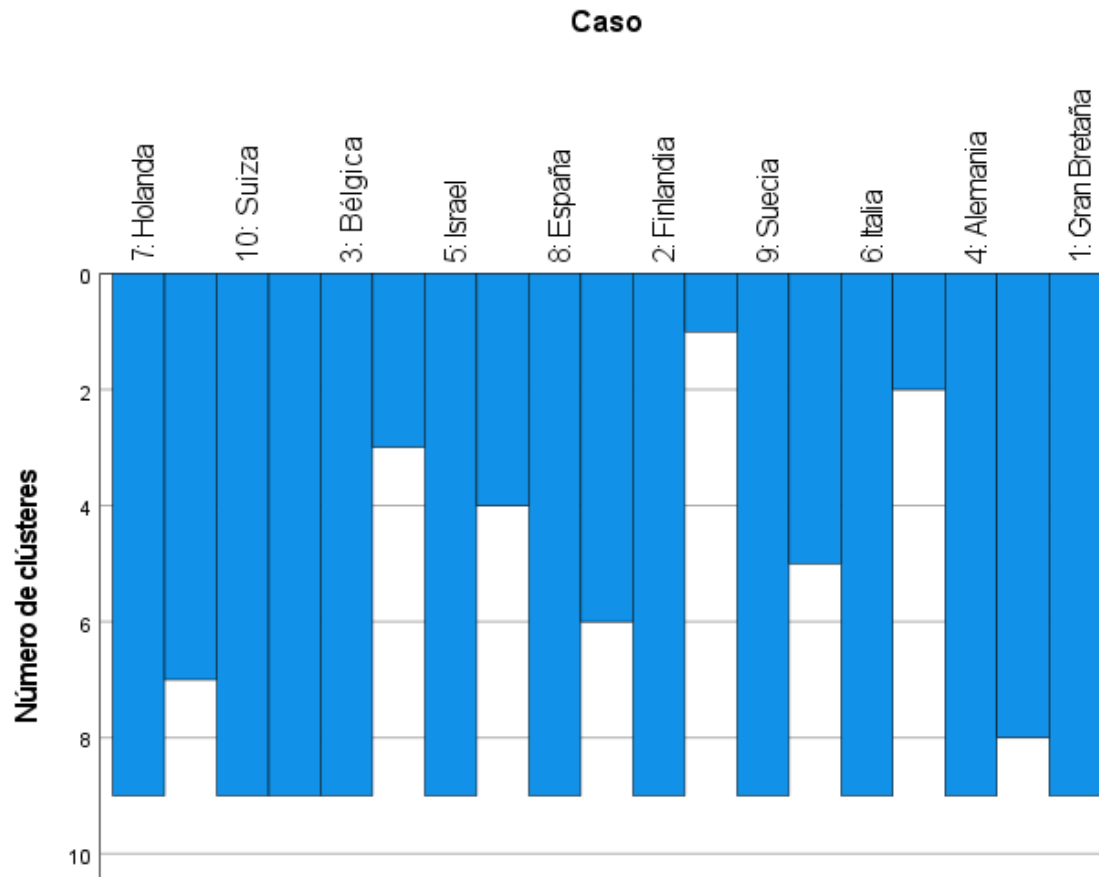
2. Dendrograma:

Dendrograma que utiliza un enlace de Ward
Combinación de clúster de distancia re-escalada



Cluster jerárquico

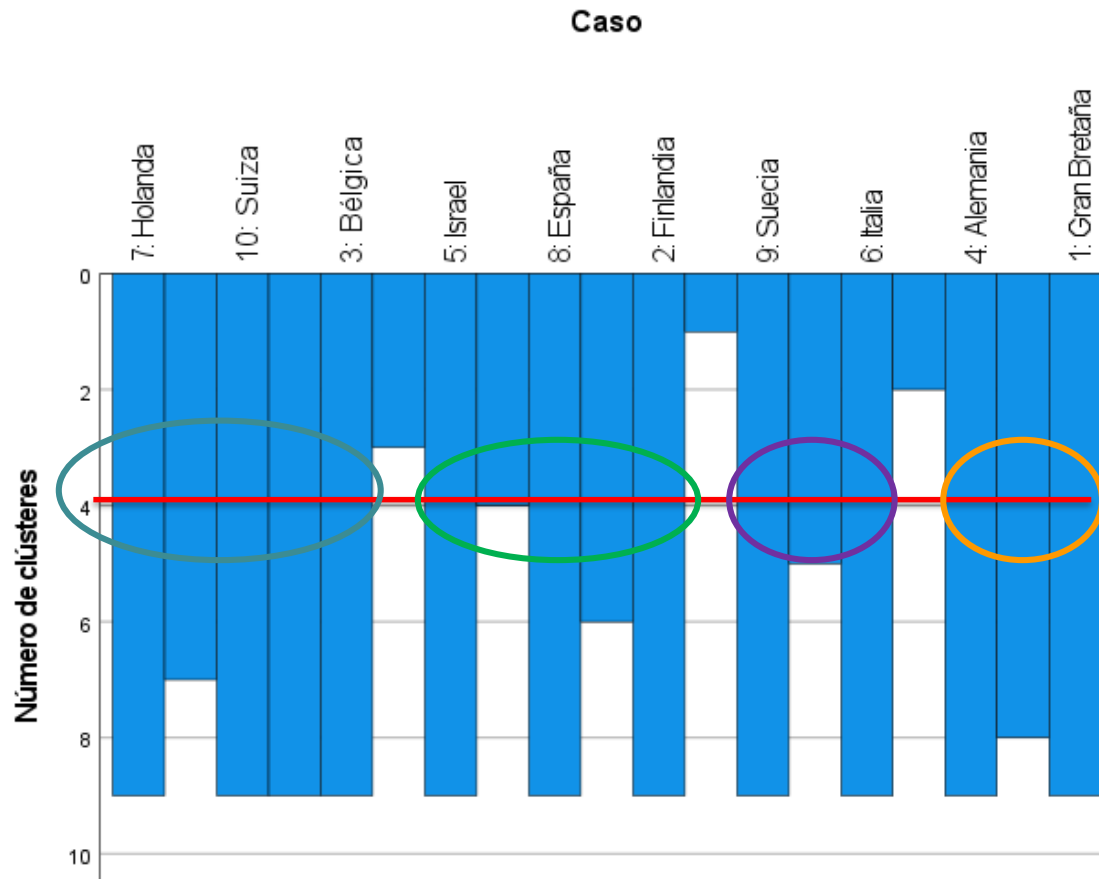
3. Diagrama de témpanos:



Los casos agrupados en cada grupo

Cluster jerárquico

3. Diagrama de témpanos:



1. **Cluster:**
2. Holanda; Suiza; Bélgica
3. **Cluster:** Israel; España; Finlandia
4. **Cluster:** Suecia; Italia
5. **Cluster:** Alemania; Gran Bretaña



EJEMPLO (TECNOLOGÍAS)

Cluster jerárquico

1. Historial de conglomeración:

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	1	4	408,570	0	0	8
2	3	10	485,290	0	0	3
3	3	7	896,875	2	0	5
4	2	5	898,860	0	0	7
5	3	8	1502,190	3	0	7
6	6	9	1687,500	0	0	9
7	2	3	2486,073	4	5	8
8	1	2	2620,722	1	7	9
9	1	6	3711,409	8	6	0

Cluster jerárquico

1. Historial de conglomeración:

Total de diferencias 45%

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	1	4	408,570	80	0	8
2	3	10	485,290	410	0	3
3	3	7	896,875	2	0	5
4	2	5	898,860	2	0	7
5	3	8	1502,190	600	0	7
6	6	9	1687,500	180	0	9
7	2	3	2486,073	4	5	8
8	1	2	2620,722	800	7	9
9	1	6	3711,409	8	6	0

Corte

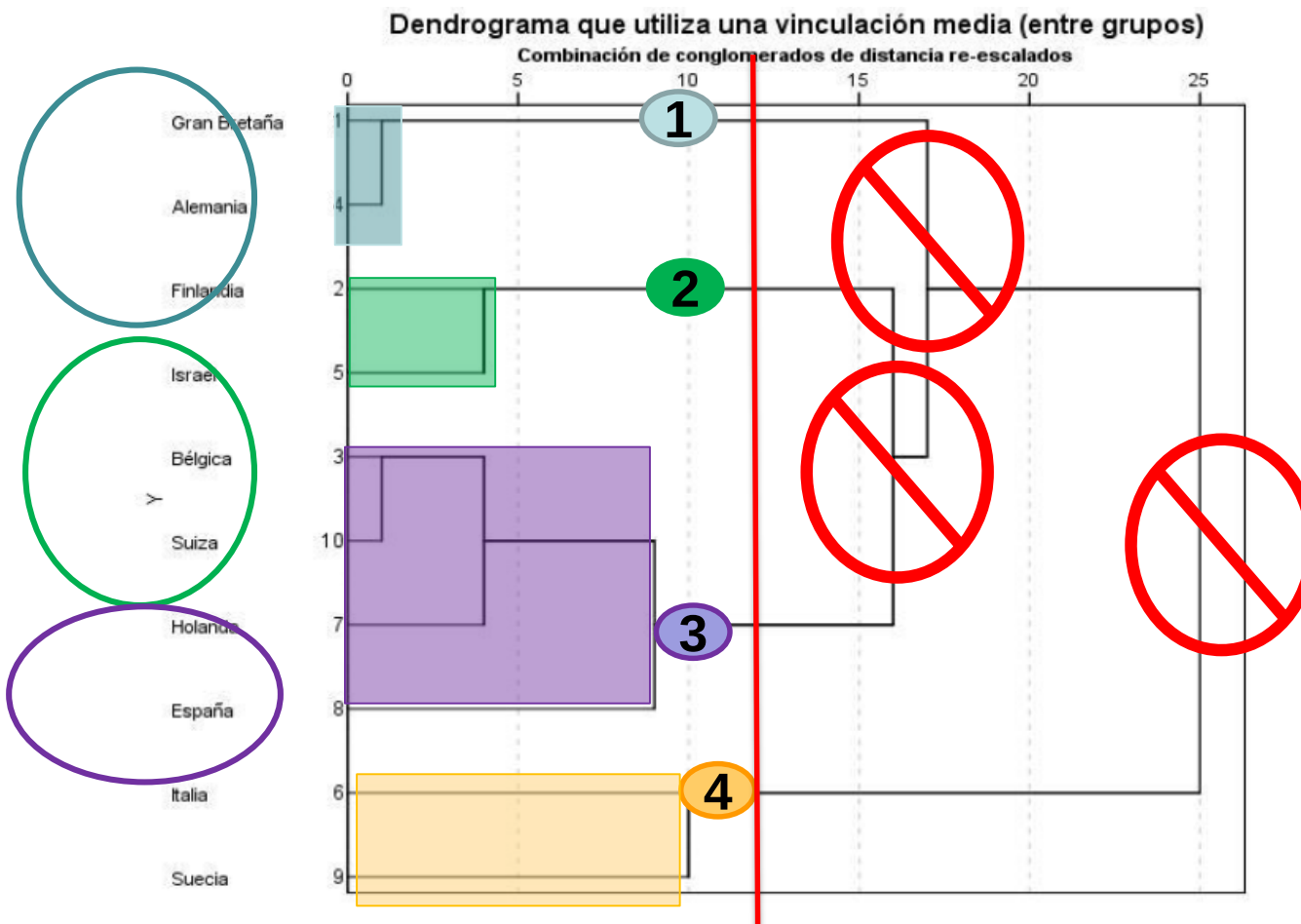
Número de grupos=Número de casos- Número de pasos permitidos

Ejemplo: 10-6=4



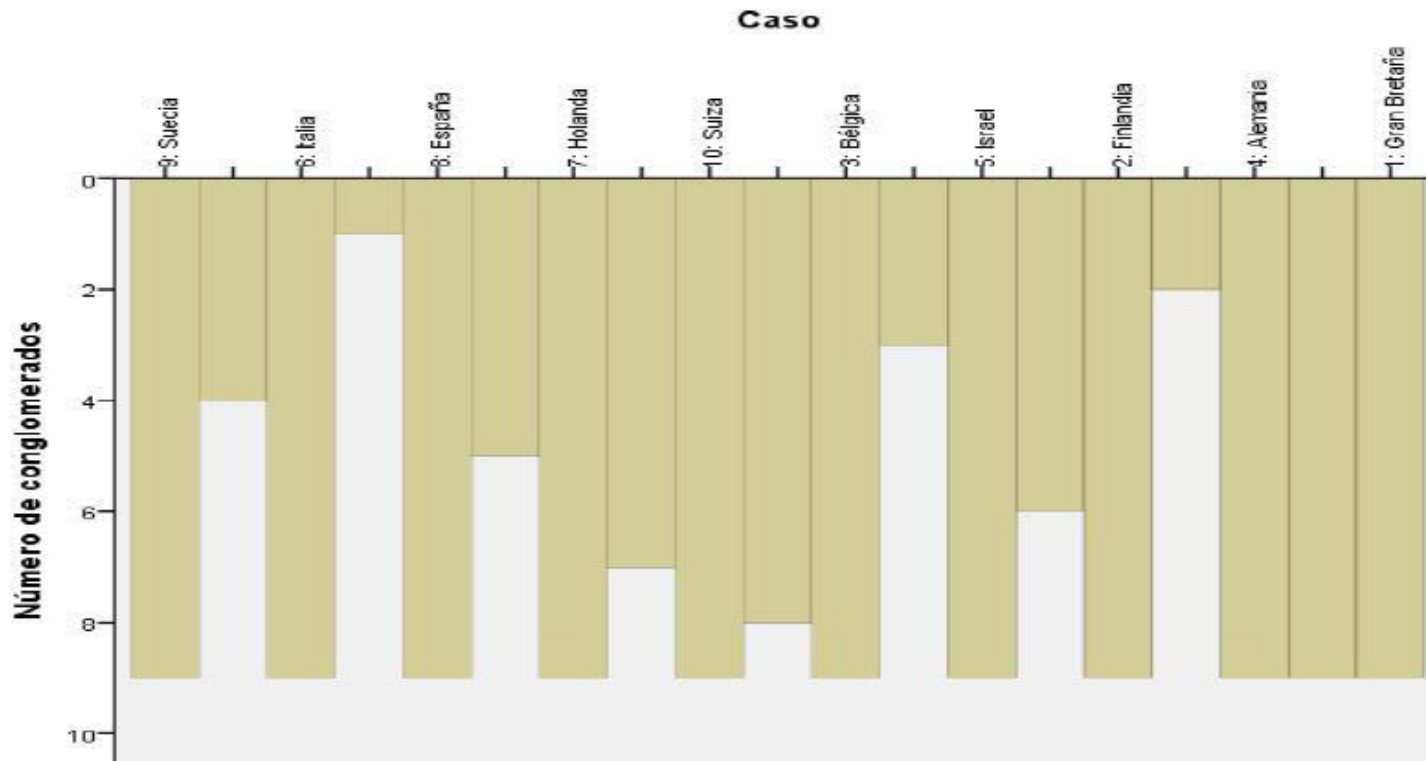
Cluster jerárquico

2. Dendrograma:



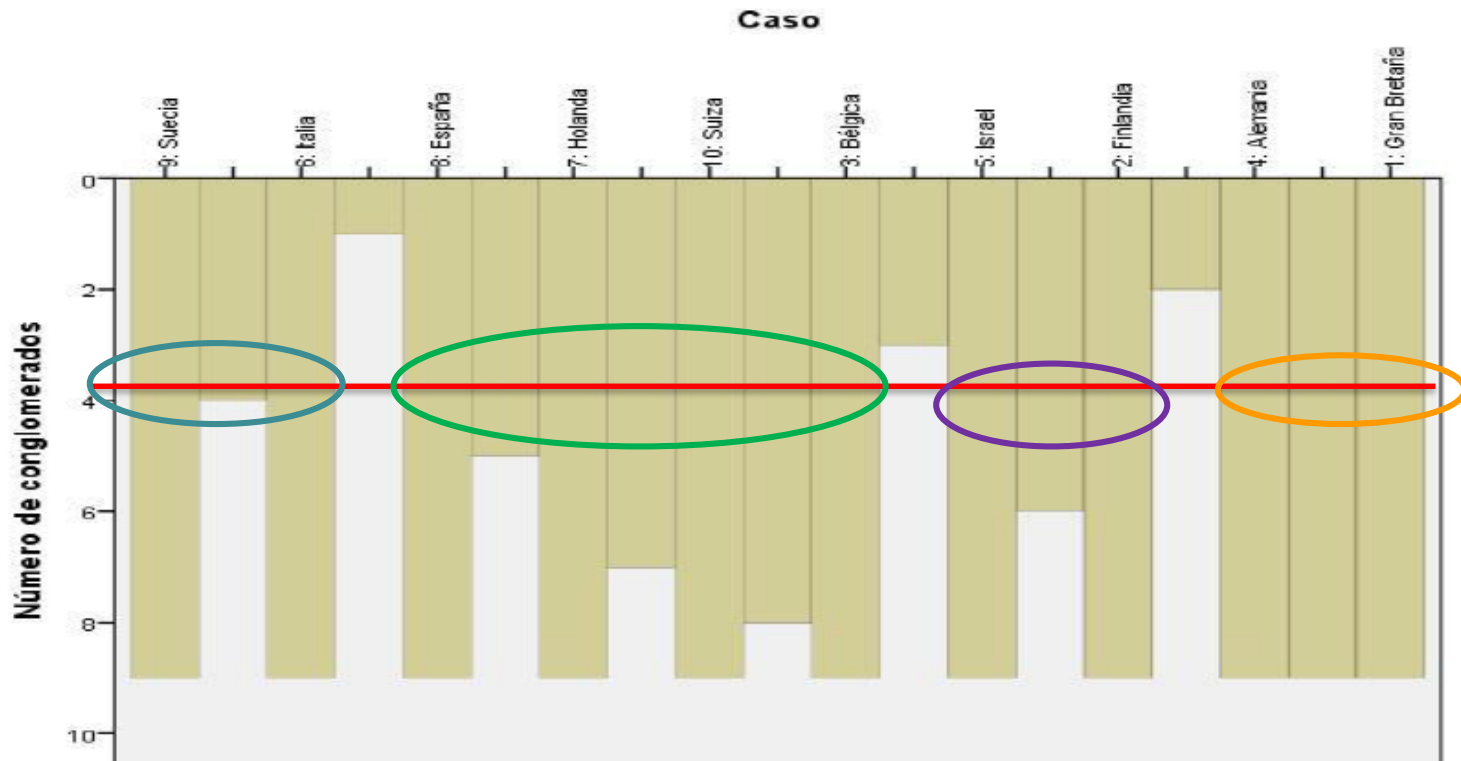
Cluster jerárquico

3. Diagrama de témpanos



Cluster jerárquico

3. Diagrama de témpanos



Cluster jerárquico

1. Historial de conglomeración:

Historial de conglomeración

Etapas	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapas siguientes
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	1	4	20,213	0	0	8
2	3	10	22,029	0	0	3
3	3	7	29,839	2	0	5
4	2	5	29,981	0	0	7
5	3	8	38,616	3	0	7
6	6	9	41,079	0	0	9
7	2	3	49,036	4	5	8
8	1	2	50,858	1	7	9
9	1	6	60,623	8	6	0

Cluster jerárquico

1. Historial de conglomeración:

Diferencias acumuladas 68%

Historial de conglomeración						
Etapas	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapas siguientes
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	1	4	20,213	2	0	8
2	3	10	22,029	7,8	0	3
3	3	7	29,839	2	0	5
4	2	5	29,981	0,1	0	7
5	3	8	38,616	9	0	7
6	6	9	41,079	2,4	0	9
7	2	3	49,036	8	5	8
8	1	2	50,858		7	9
9	1	6	60,623	8	6	0

Corte

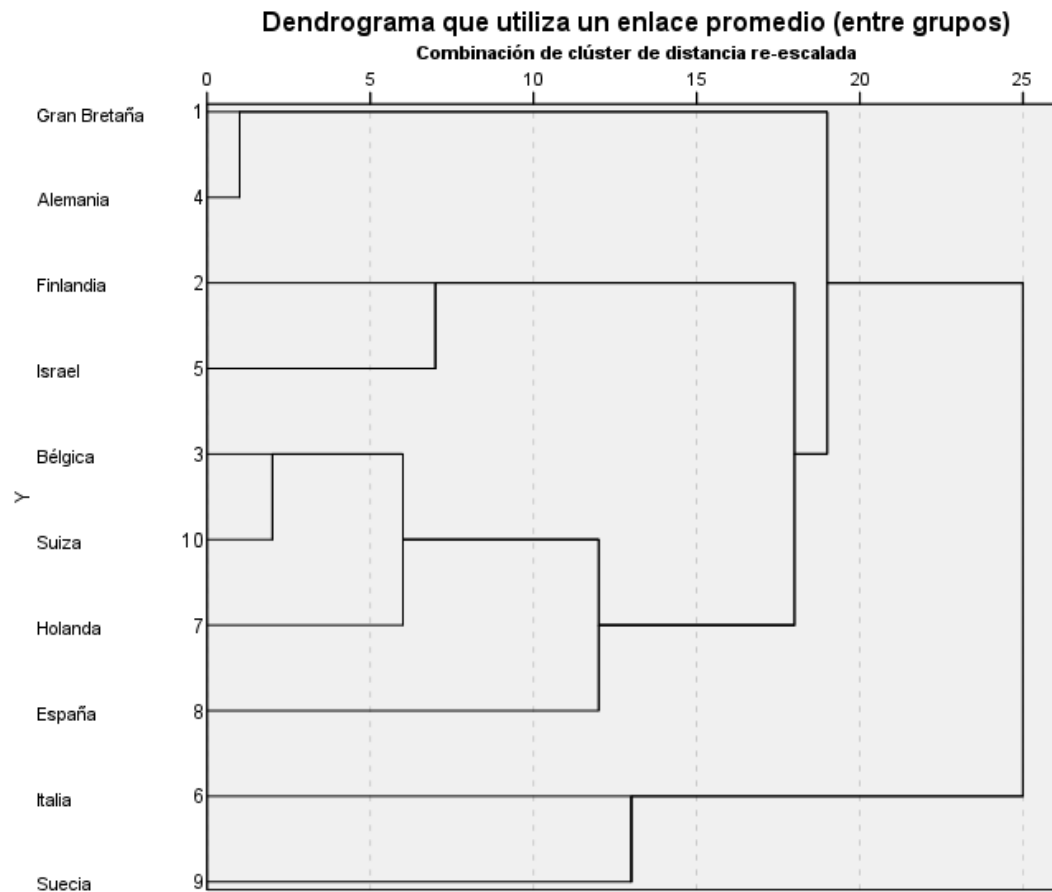
Número de grupos=Número de casos- Número de pasos permitidos

Ejemplo: $10-6=4$



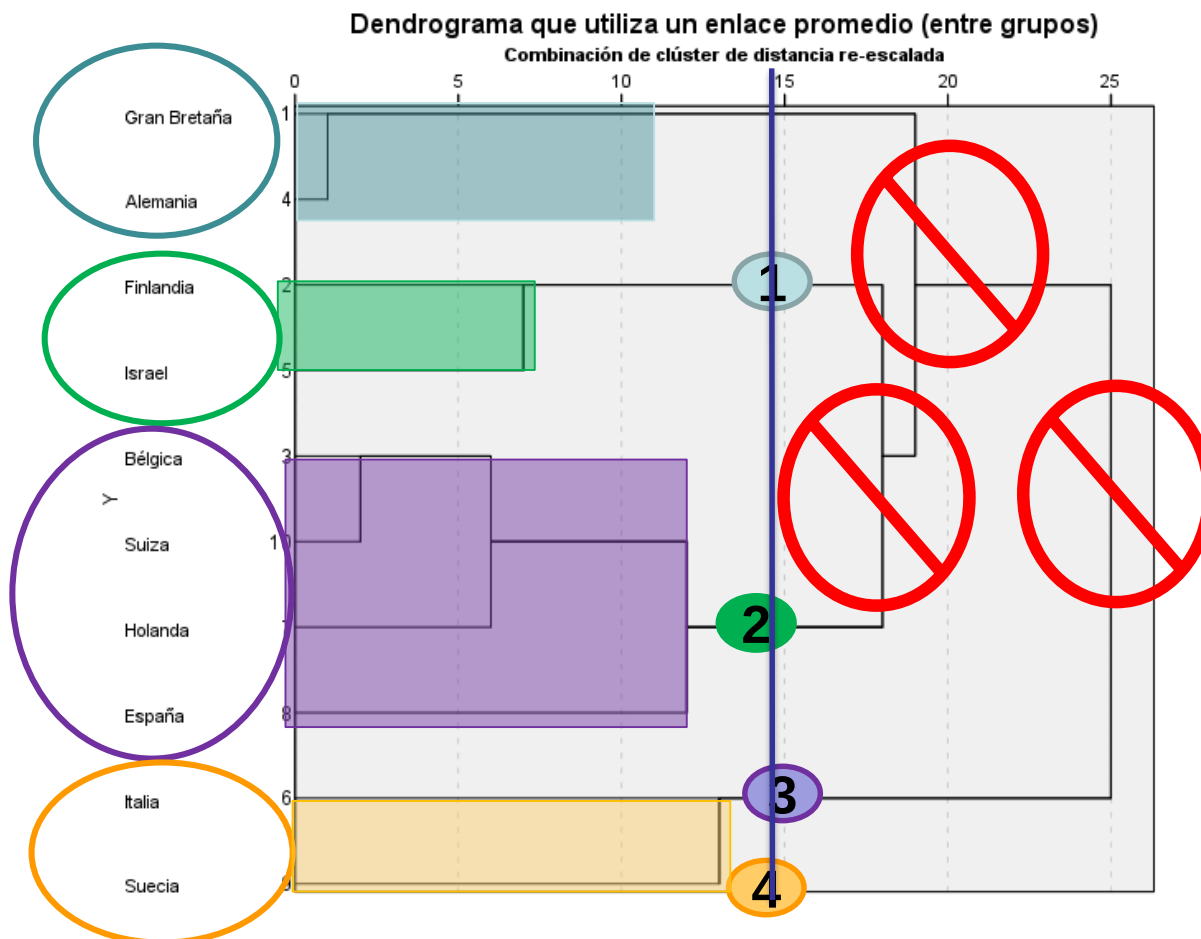
Cluster jerárquico

2. Dendograma:



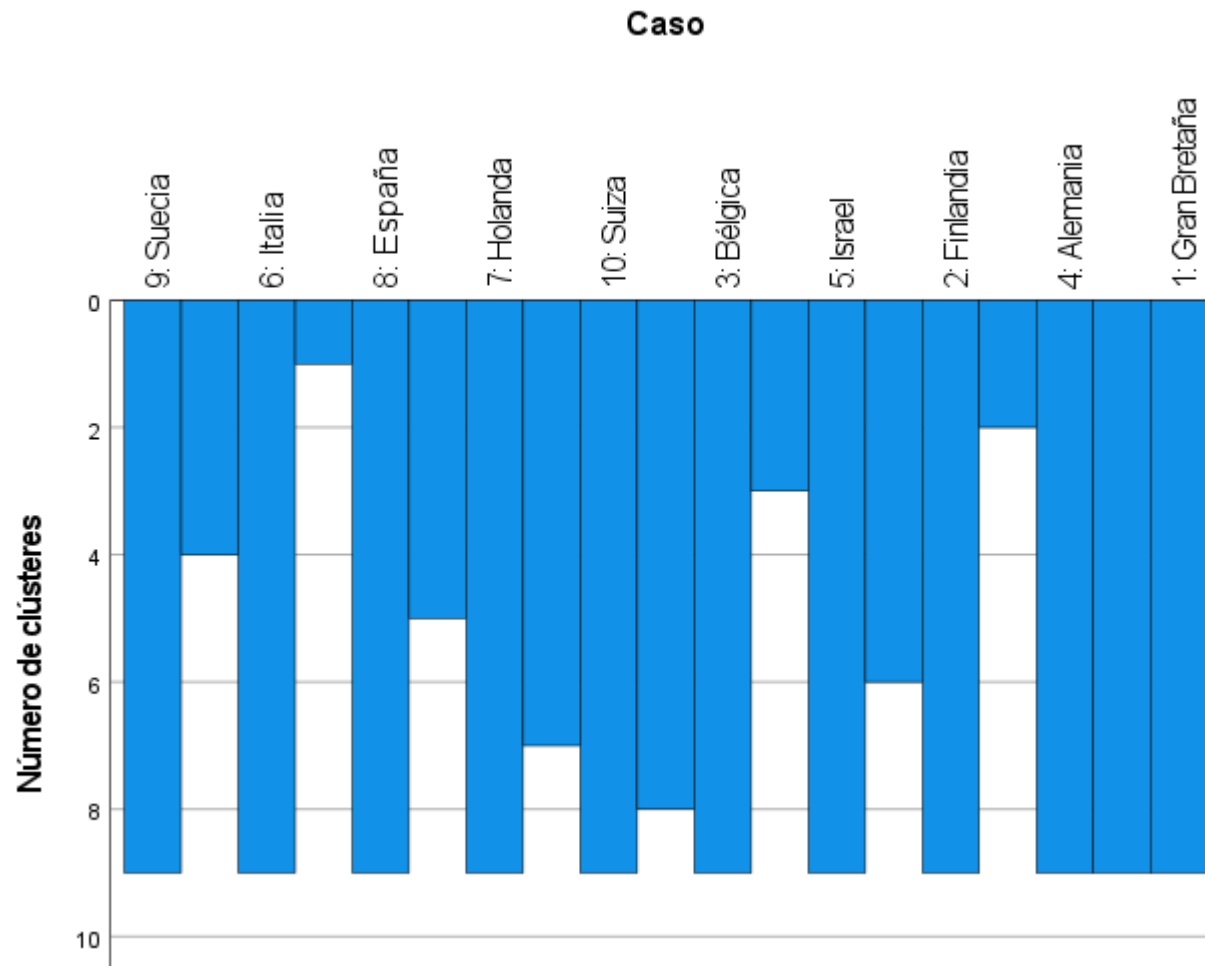
Cluster jerárquico

2. Dendrograma:



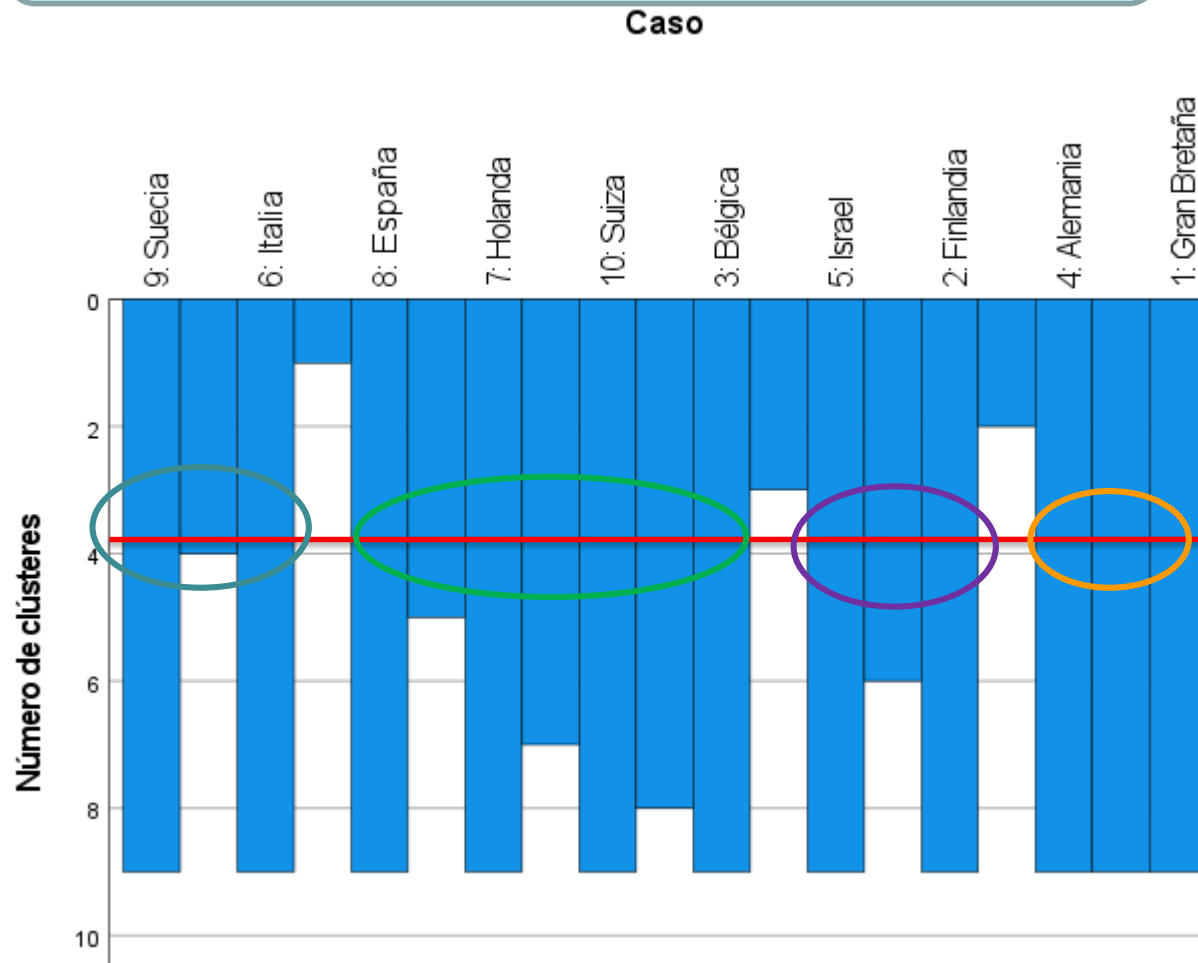
Cluster jerárquico

3. Diagrama de témpanos



Cluster jerárquico

3. Diagrama de témpanos:



1. **Cluster:** Suecia;
Italia

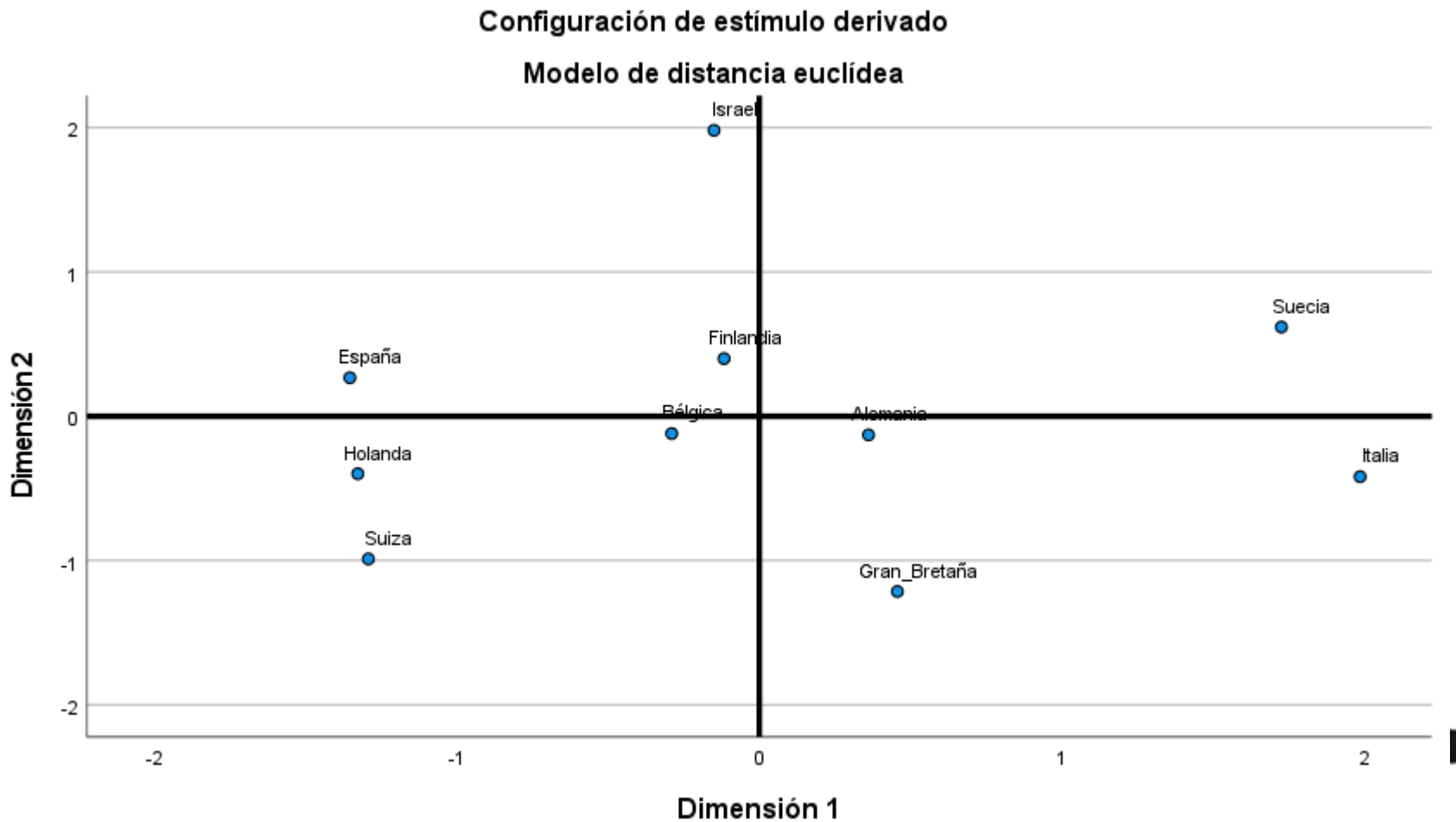
2. **Cluster:** España;
Holanda; Suiza;
Bélgica;

3. **Cluster:** Israel,
Finlandia;

4. **Cluster:** Alemania;
Gran Bretaña

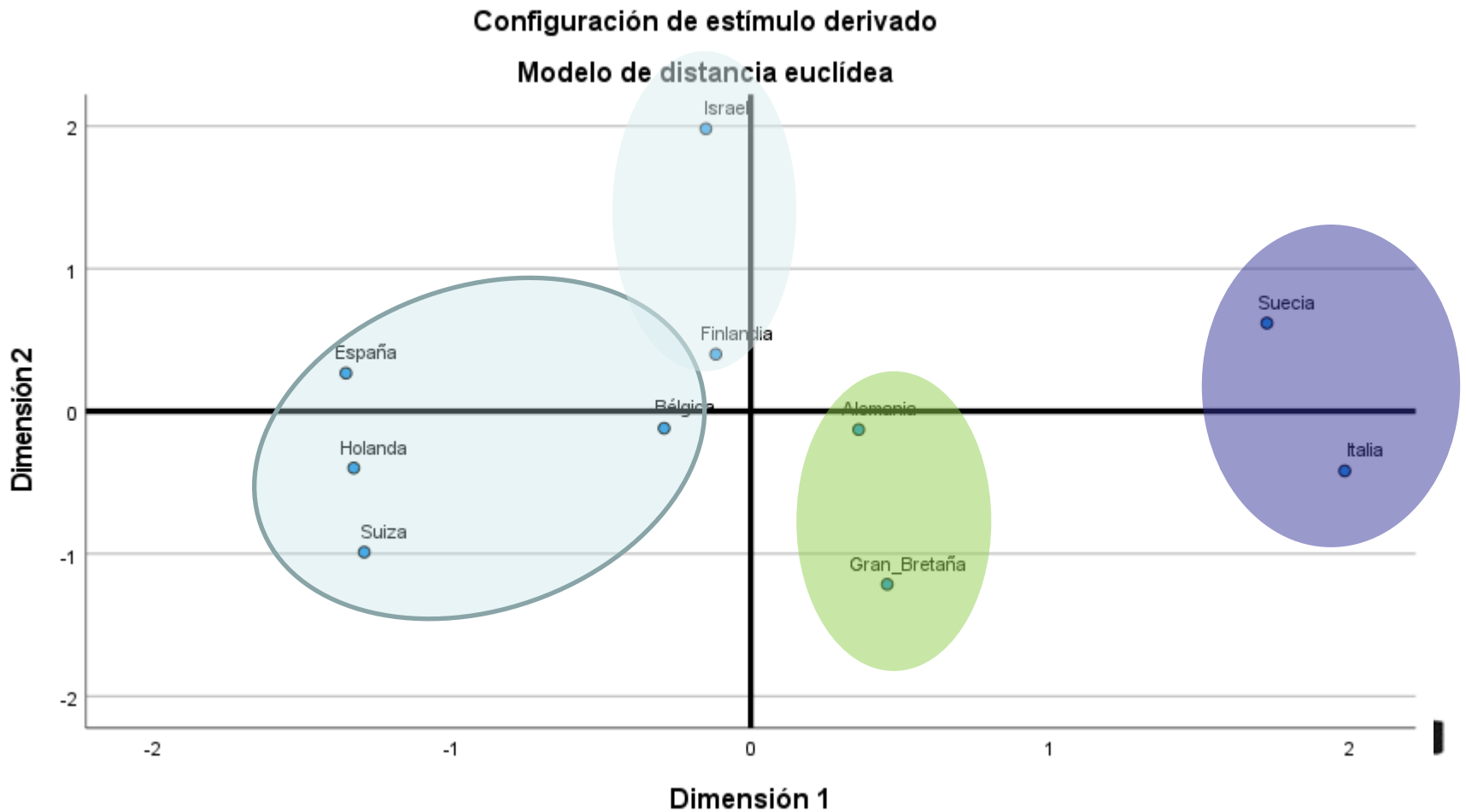
Cluster jerárquico

Por qué varía?



Cluster jerárquico

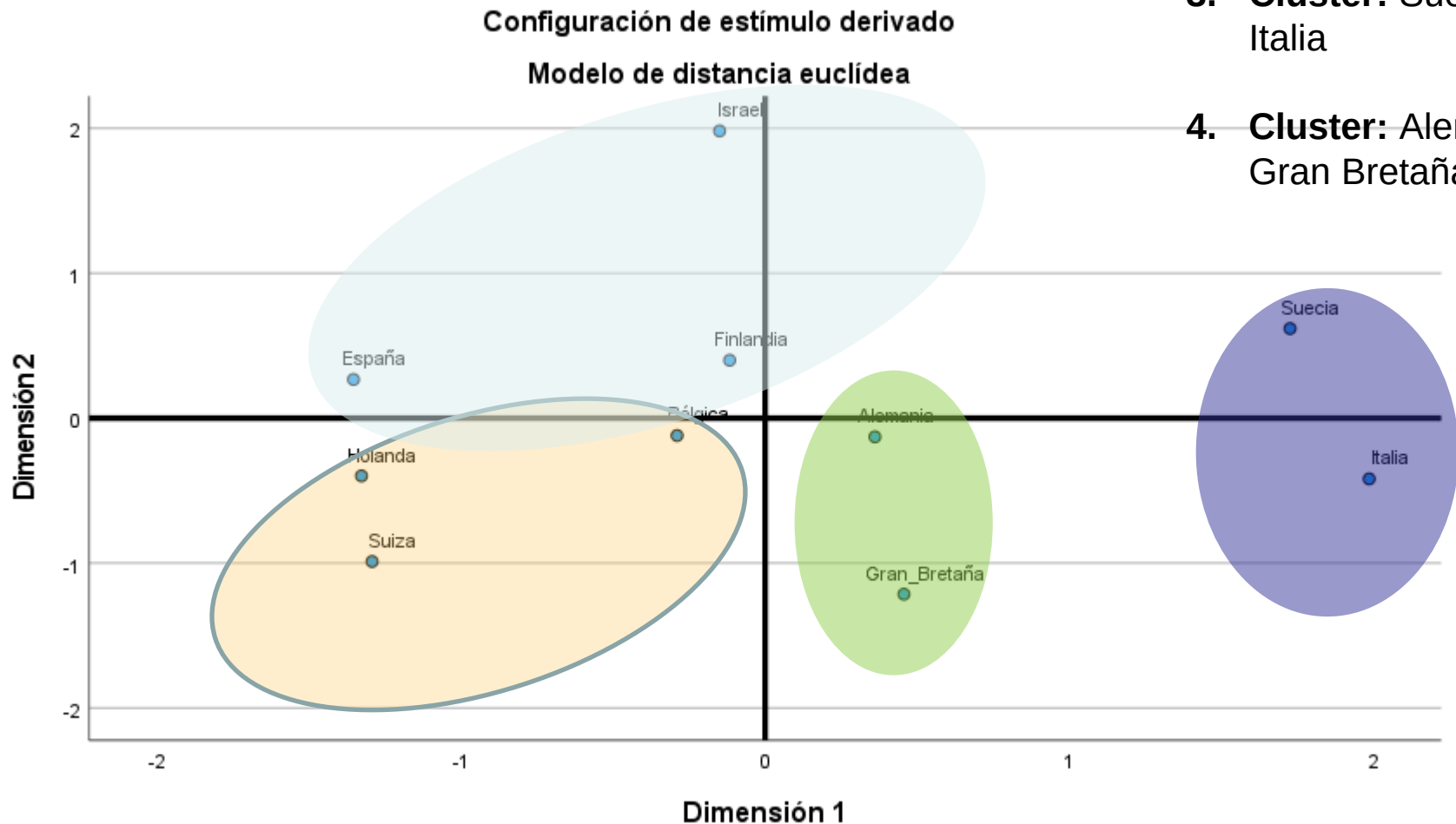
Por qué varía?



Cluster jerárquico

Por qué varía?

1. **Cluster:** Holanda; Suiza; Bélgica
2. **Cluster:** Israel; España; Finlandia
3. **Cluster:** Suecia; Italia
4. **Cluster:** Alemania; Gran Bretaña





EJEMPLO (TITULACIONES)

Cluster jerárquico

1. Historial de conglomeración:

Historial de conglomeración

Etapa	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapa siguiente
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	5	10	33,980	0	0	2
2	4	5	100,467	0	1	3
3	1	4	217,575	0	2	5
4	7	8	355,240	0	0	8
5	1	9	562,365	3	0	6
6	1	2	912,718	5	0	7
7	1	3	1893,971	6	0	9
8	6	7	3313,899	0	4	9
9	1	6	5357,341	7	8	0

Cluster jerárquico

1. Historial de conglomeración:

Total diferencias acumuladas **69%**

Historial de conglomeración

Etapa	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapa siguiente
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	5	10	33,980	66	0	2
2	4	5	100,467	117	1	3
3	1	4	217,575	0	2	5
4	7	8	355,240	140	0	8
5	1	9	562,365	200	0	6
6	1	2	912,718	350	0	7
7	1	3	1893,971	980	0	9
8	6	7	3313,899	1400	4	9
9	1	6	5357,341	2000	8	0

Corte?

Número de grupos=Número de casos- Número de pasos permitidos

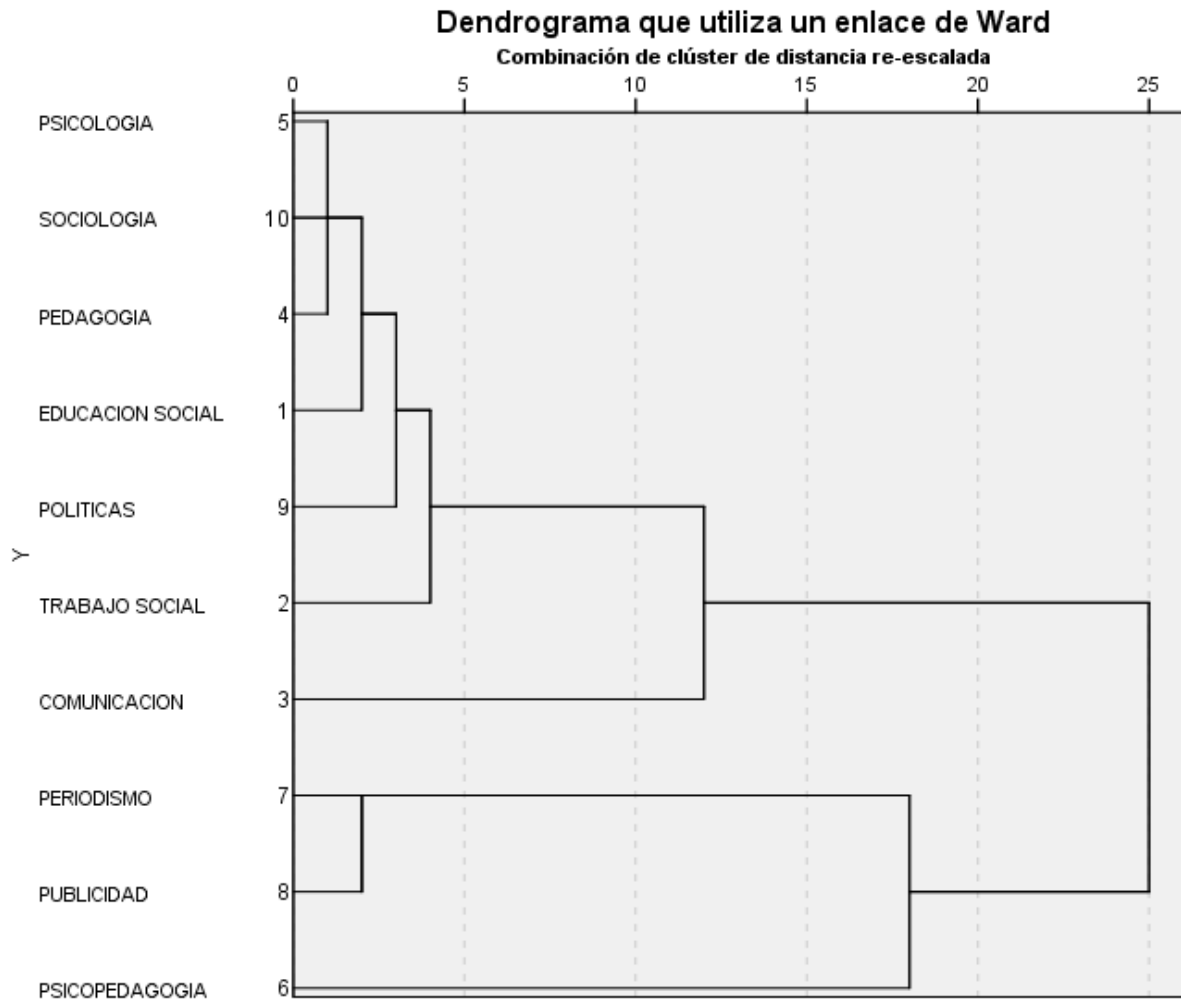
$$H1: 10-6=4$$

$$H2: 10-7=3$$



Cluster jerárquico

2. Dendograma:

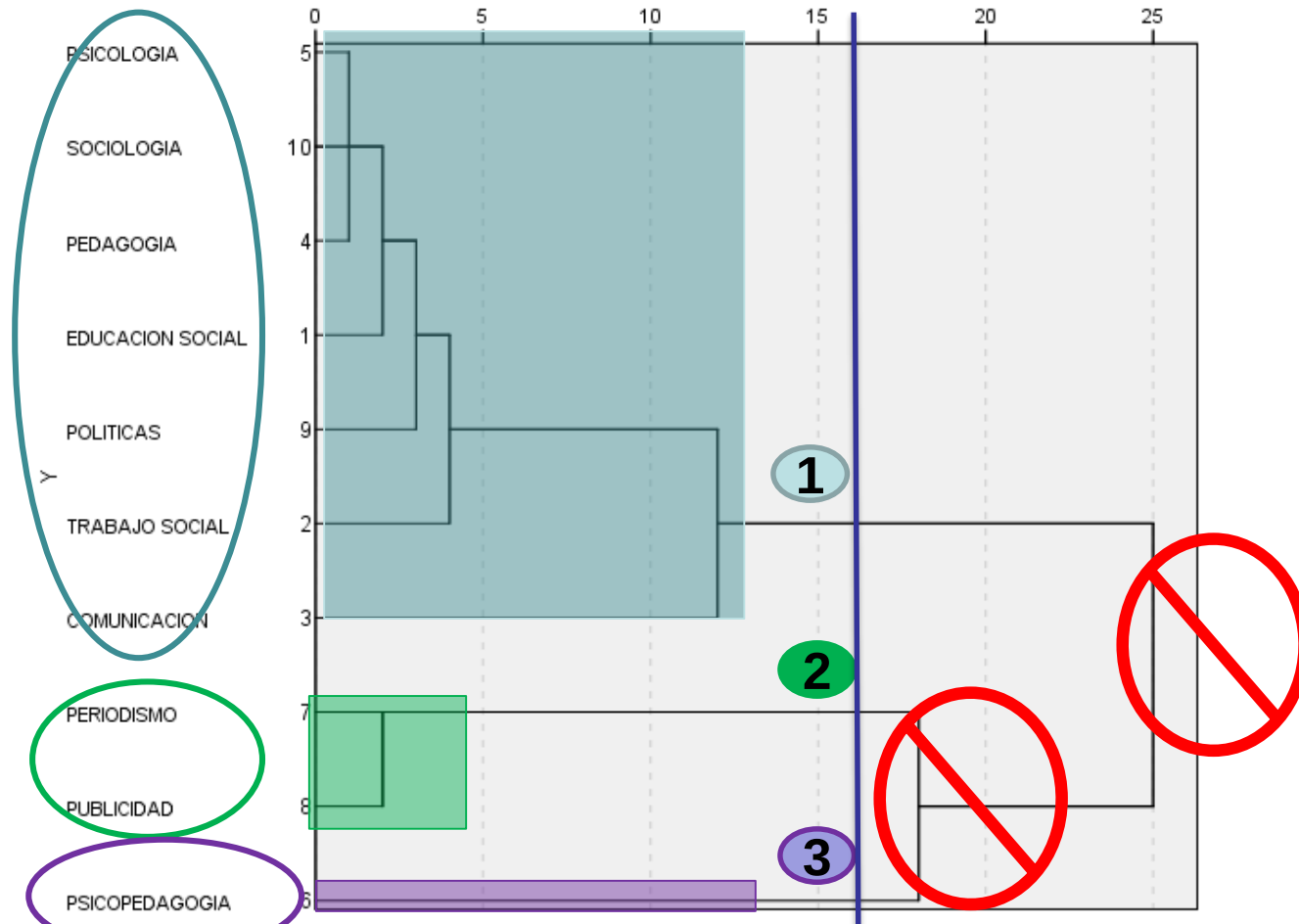


Cluster jerárquico

2. Dendograma:

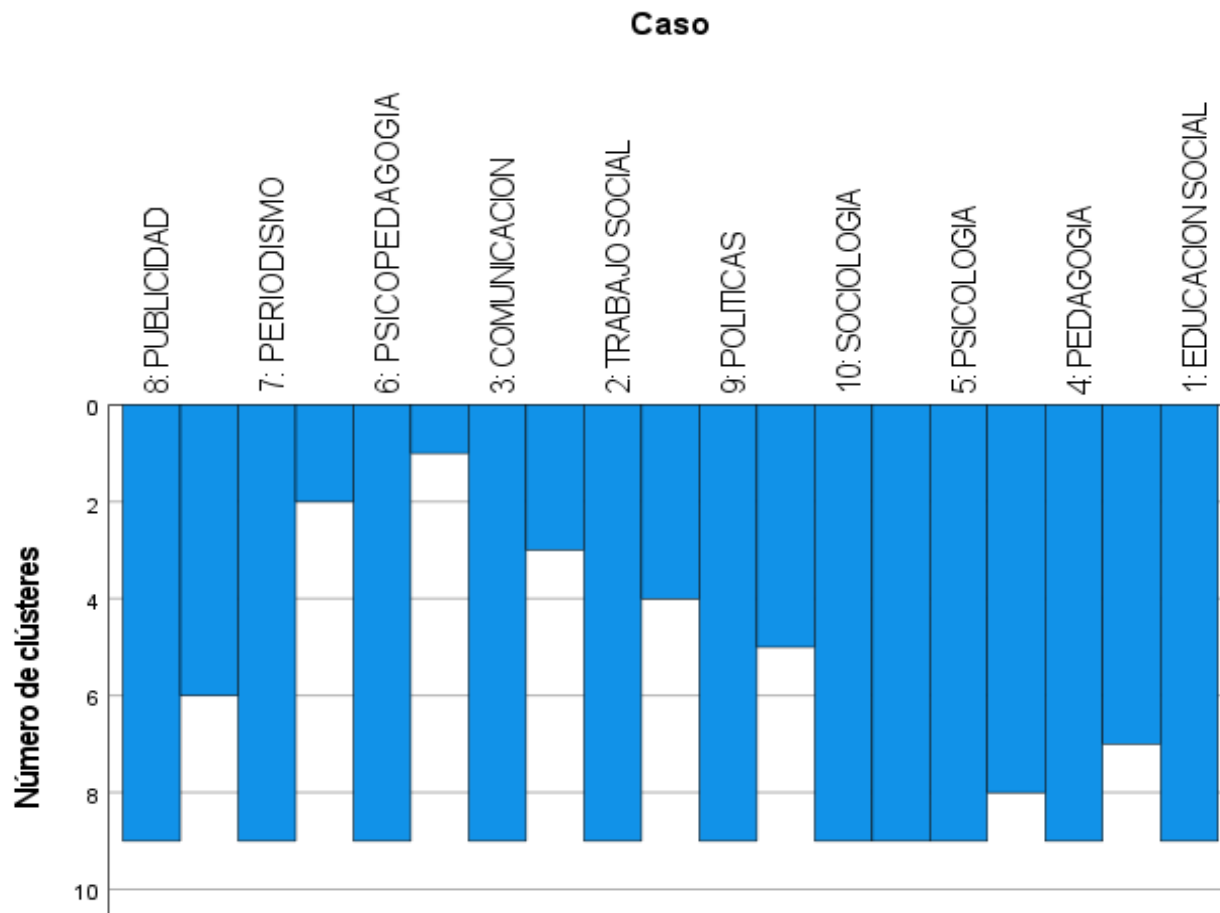
Dendograma que utiliza un enlace de Ward

Combinación de clúster de distancia re-escalada



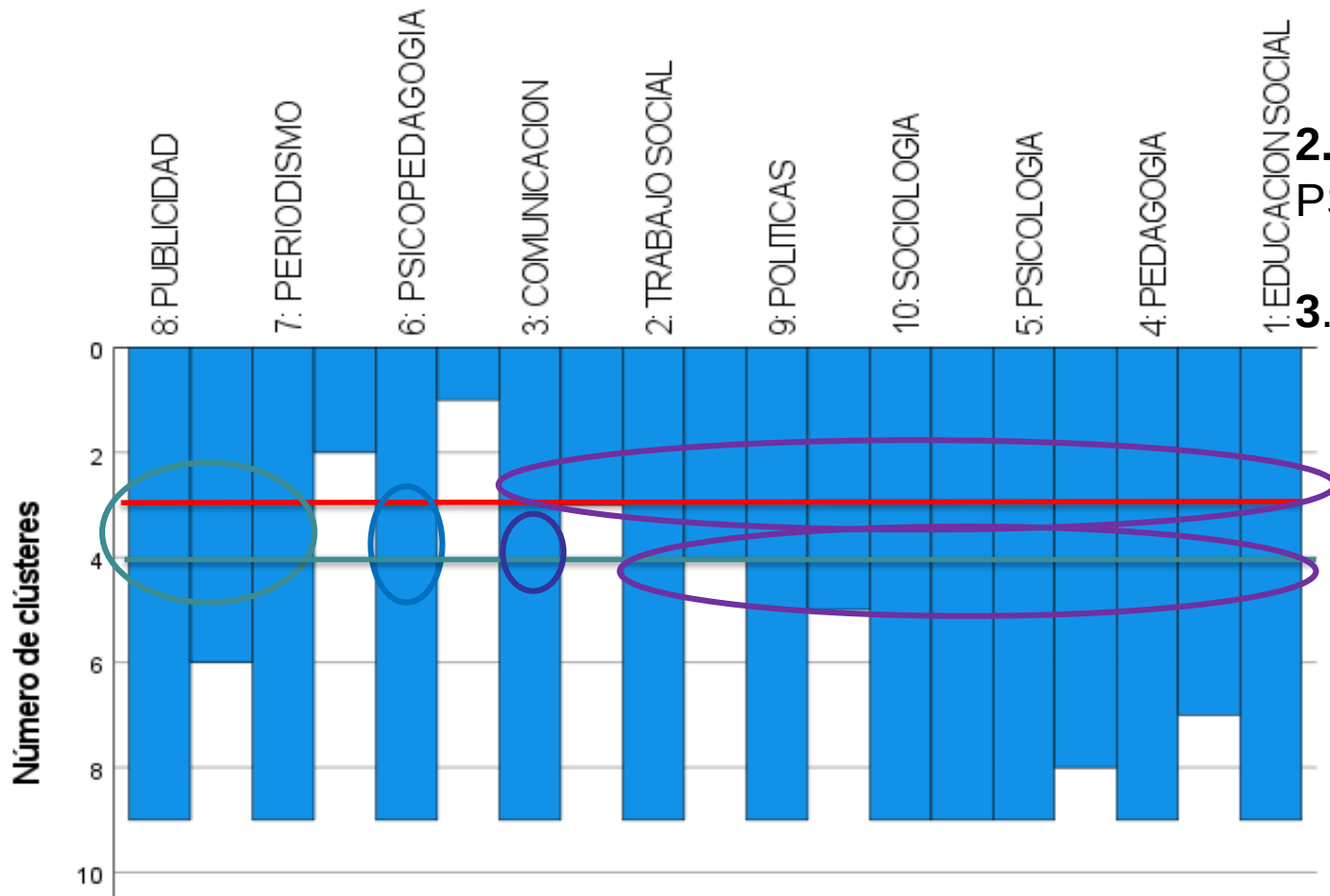
Cluster jerárquico

3. Diagrama de témpanos



Cluster jerárquico

3. Diagrama de témpanos



1. Cluster: PUB, PERI

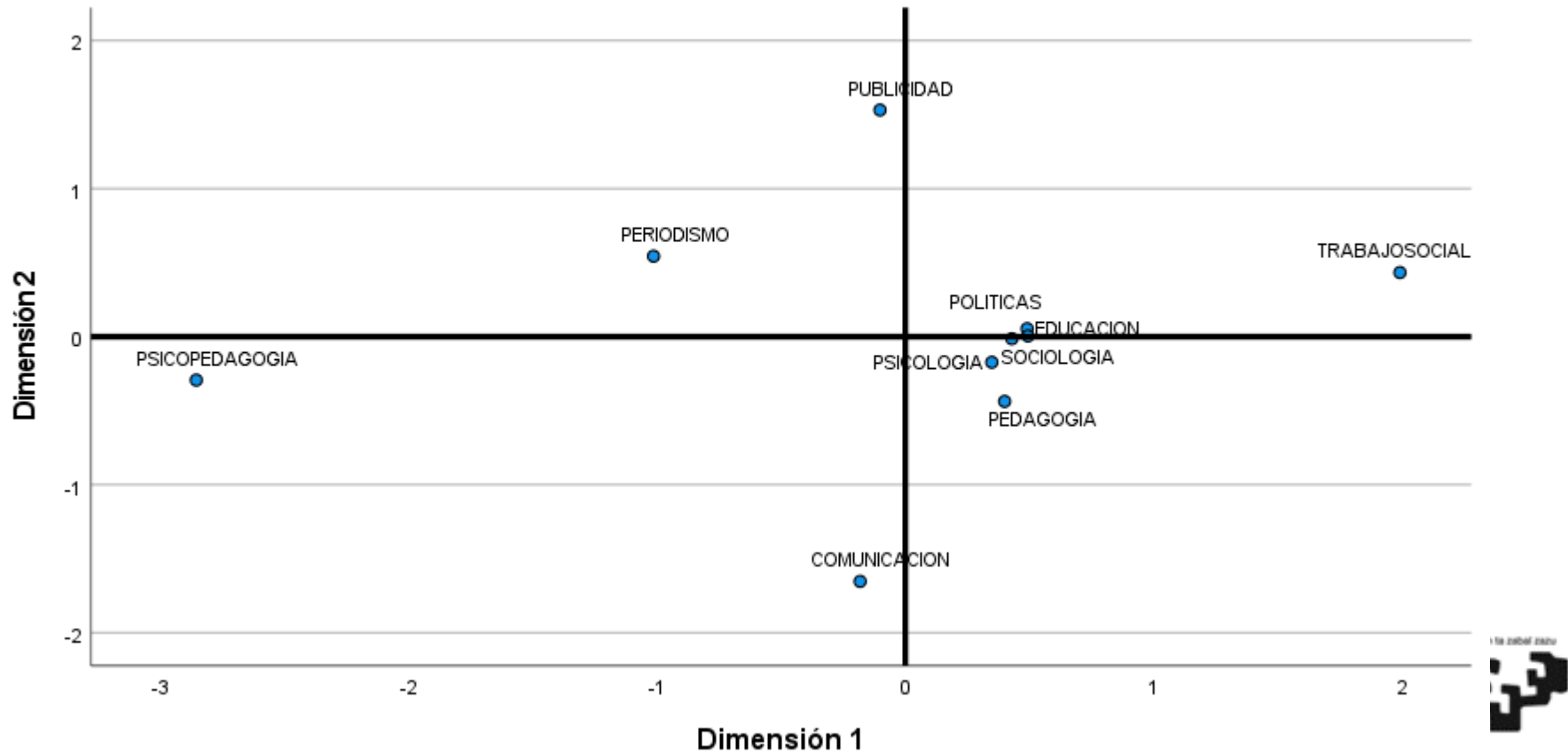
2. Cluster: PSICOPED

3. Cluster: COM, TS, POL, SOCIO, PS, PED, ES

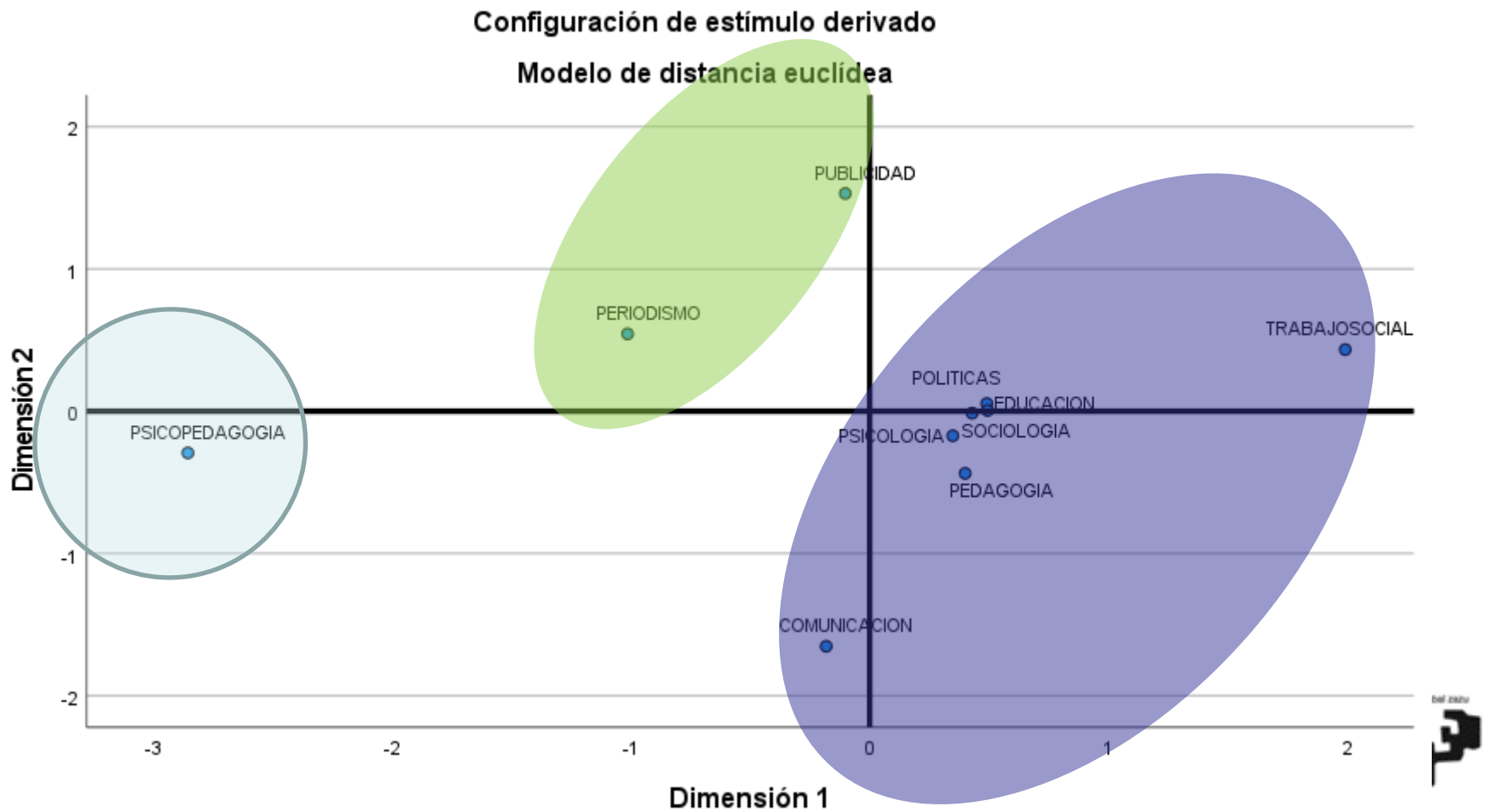
Cluster jerárquico

Configuración de estímulo derivado

Modelo de distancia euclídea



Cluster jerárquico



Cluster jerárquico

